

The Contribution of Retrotransposons to the Transcriptomes of Murine Somatic Cells

Joseph Michael Gardner
Girton College

Supervisor: Prof. Anne Ferguson-Smith

This dissertation is submitted
for the degree of Doctor of Philosophy

September 2017

Preface

I declare that

- This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.
- It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.
- It does not exceed the prescribed word limit for the relevant Degree Committee.

Summary

The Contribution of Retrotransposons to the Transcriptomes of Murine Somatic Cells

Joseph Michael Gardner

Retrotransposons comprise approximately 40% of the mouse genome. Once thought to be useless “junk” DNA, there is growing evidence that retrotransposons play crucial roles in genome evolution and gene regulation, and contribute to the transcriptome. Several studies have found functional retrotransposon transcripts in the germline and during early development, but less is known about retrotransposon transcription in adult somatic cells. Retrotransposons are also responsible for generating gene copies in mammalian genomes (retrocopies), and there are several examples of retrocopies evolving into new genes, or being transcribed as non-coding RNA. Using computational approaches, I analyse RNA-seq data to assess the contribution of retrotransposons and retrocopies to the transcriptomes of adult mouse somatic cells, using purified naive B and T lymphocytes. First, I describe the transcriptomes generated using high-quality total RNA-seq data. Second, I quantify and characterise the retrotransposon content of these transcriptomes. Finally, I identify retrocopy transcripts and assess their relationship with

the genes from which they originate. I found widespread inclusion of retrotransposons in somatic cell transcriptomes. These transcripts form distinct clusters based on retrotransposon sequence, with endogenous retroviruses being particularly prevalent in retrotransposon-rich transcripts. While these clusters are consistent between cell types, the individual retrotransposons transcribed show cell-type specificity. I also find evidence that retrotransposons may facilitate gene regulation by antisense transcripts. I demonstrate that a subset of retrocopies is transcribed, and the vast majority of these form RNA complementary to their parent mRNA, with high sequence identity. Using differential expression and proteome analysis, I present evidence for post-transcriptional regulation of parent transcripts by retrocopy RNA, possibly through stabilisation of the parent RNA. I also find that while retrocopy expression is not necessarily shared between cell types or mouse strains, certain parent transcripts tend to have an expressed retrocopy in multiple contexts. Overall, this thesis presents evidence of an important role for retrotransposons and retrocopies in the adult somatic transcriptome, and sets the stage for further investigation to experimentally elucidate the functions of these transcripts.

Acknowledgements

There are several people I would like to thank.

First, my supervisor, Anne Ferguson-Smith, for her advice, support, patience and trust. Second, to the members of the AFS lab, past and present - the best group of people I ever baked for. Particular thanks go to Anastasiya, Carol, Celia, Hui, Jenni, Lisa, Marcela, Nic, Rahia, and Tessa, for their support, help, and friendship. I would also like to thank Russell Hamilton and Sudhakaran Prabakaran for their advice, and Claire Wilson, for her help on many matters non-scientific. Thanks to my advisor Aylwyn Scally, for his useful feedback on this project as it evolved, and to my examiners David Kent and Colin Semple. Thanks to Julia Gog for being an excellent lecturer and showing me the way from mathematics to biology. Thank you to the BBSRC for funding this PhD.

I would also like to thank my friends, for keeping me grounded in real life: Zelandar, Stephen, Fred, Susanne, Sophia, Giovanni, Vincenzo; and, in particular, my housemates Matthew and Nicolas - it was definitely a home, not just a house. To my wife, Matilde: I couldn't have done this without you. To my brothers, Rob and Alex. And finally, thank you to my parents, for everything.

Related Publications

Identification, characterization and heritability of murine metastable epialleles: implications for non-genetic inheritance. Kazachenka, Bertozzi, *et al.*, *Cell* 2018

Contents

1	Introduction	37
1.1	Long Non-Coding RNA	39
1.2	Retrotransposons in the Mouse Genome	43
1.2.1	Introduction to the Repetitive Genome	43
1.2.1.1	Retrotransposons	45
1.2.1.2	LINEs	48
1.2.1.3	SINEs	51
1.2.1.4	ERVs	52
1.2.2	Silencing of Retrotransposons	54
1.2.2.1	DNA Methylation	54
1.2.2.2	Histone Modifications	56
1.2.2.3	Piwi-interacting RNA in the Germline	58
1.2.3	The Role of Retrotransposons in Genome Evolution and Function	61
1.2.3.1	Retrotransposons Create Regulatory Elements . . .	62
1.2.3.2	Retrotransposons Cause Genomic Rearrangements	65
1.2.3.3	Retrotransposons Alter the Epigenome	67

1.2.4	Transcription of Retrotransposons	68
1.2.4.1	Retrotransposon Transcription in Pluripotent Cells	69
1.2.4.2	Retrotransposon Transcription in Adult Cells . . .	69
1.2.5	Retrotransposon Outlook	72
1.3	Gene Retrocopies	73
1.3.1	Retrocopy Origins in Mammals	73
1.3.2	Contribution of Retrocopies to Genome Evolution	77
1.3.3	Retrocopy Expression and the Germline	78
1.3.4	Non-coding Transcription of Retrocopies	79
1.3.4.1	Targeting of Epigenetic Modifications	80
1.3.4.2	Retrocopy LncRNAs as Micro RNA Decoys	80
1.3.4.3	Retrocopy Antisense LncRNAs	81
1.3.4.4	Retrocopy Short RNAs	82
1.3.5	Retrocopy Regulation	83
1.3.6	Retrocopy Outlook	83
1.4	RNA Sequencing	84
1.4.1	Summary of Experimental Procedure	84
1.4.1.1	RNA Preparation	85
1.4.1.2	Library Preparation	87
1.4.1.3	Sequencing	88
1.4.2	Bioinformatic Analysis	89
1.4.2.1	RNA-seq and Repeats	91
1.5	Motivation and Open Questions	92
1.5.1	Aims	93

2	Datasets	95
2.1	BLUEPRINT Datasets	95
2.1.1	Mouse Strains	96
2.1.2	Sexes	96
2.1.3	Cell Types	96
2.1.4	RNA Sequencing	98
2.1.5	Whole Genome Bisulphite Sequencing	98
2.1.6	Chromatin Immunoprecipitation Sequencing	99
2.2	ENCODE Data	100
2.3	Proteomes	100
2.4	Reference Genomes and Annotations	101
2.4.1	Reference Genomes	101
2.4.2	Reference Transcriptome	101
2.4.3	Repeat Annotation	101
2.4.4	Retrocopy Annotation	105
3	Methods	107
3.1	Transcriptome Reconstruction Pipeline	107
3.1.1	Comparison of Available Tools	108
3.1.1.1	Quality Control	108
3.1.1.2	Alignment and rRNA Removal	109
3.1.1.3	Transcript Reconstruction	110
3.1.1.4	Abundance Estimation	110
3.1.2	Final Pipeline	111
3.2	Differential Expression Analysis	112

3.2.1	Differential Expression of Reconstructed Transcripts	112
3.2.2	Differential Expression of Ensembl Transcripts	113
3.3	Comparing Reconstructed and Ensembl Transcripts	114
3.4	Ribosomal RNA in CAST	115
3.5	Epigenetic State Visualisation	116
3.6	Proteome Normalisation	118
3.7	Transcript Retrotransposon Content	119
3.8	Retrotransposon Content Visualisation	121
3.9	Cluster Comparison	124
3.10	Expression Correlation Distributions	129
3.11	Venn Diagrams and Statistical Analysis	129
3.12	Retrocopy Transcripts	131
3.13	Retrocopy Conservation	131
3.14	Retrocopies in CAST	132
3.15	Computing Environments and Online Resources	136
4	Mouse Lymphocyte Transcriptomes	139
4.1	Alignments	139
4.2	Reconstructed Transcriptomes	146
5	Retrotransposon Transcription in Lymphocytes	155
5.1	Quantifying Retrotransposon Transcription	156
5.2	Effect of RT transcripts on gene expression	162
5.3	Cell-type Specificity	172
5.4	Summary	177

6	Retrocopy Transcription in Mouse Lymphocytes	179
6.1	Retrocopies are Expressed in Mouse Lymphocytes	180
6.2	Retrocopy expression is shared across lineages	185
6.3	Expressed Retrocopies Produce RNA Complementary to Their Parent	188
6.4	Expressed Retrocopies Have Higher Sequence Identity with Their Parents	191
6.5	Retrocopy Transcription and Retrotransposon Indels	195
6.6	Retrocopy Expression May Affect Parent mRNA Levels	200
6.7	Retrocopy Transcription Does Not Affect Epigenetic State of the Parent Locus	207
6.8	Retrocopy Expression May Affect Protein Abundance	208
6.9	Conservation of BL6 Expressed Retrocopies in CAST	211
6.10	Summary	216
7	Discussion	219
7.1	Retrotransposon Transcription	219
7.1.1	Summary of Results	219
7.1.2	Comparison to Existing Literature	220
7.1.3	Conclusions and Future Work	223
7.2	Retrocopy Transcription	226
7.2.1	Summary of Results	226
7.2.2	Comparison to Existing Literature	227
7.2.3	Conclusions and Future Work	229
7.3	Final Remarks	232

List of Figures

1.1	The number of lncRNAs found in human and mouse in different transcriptomics studies. Figure based on Table 1 in [18], using data from [13–15, 19–23]	41
1.2	The four broad classes of repeats.	44
1.3	The repetitive DNA content of the mm10 mouse and hg38 human reference genomes, as identified by RepeatMasker [8]. LINE: long interspersed element. ERV: endogeneous retrovirus. SINE: short interspersed element. SVA: SINE VNTR <i>Alu</i> ; a primate-specific retrotransposon composed of a SINE, a variable number tandem repeat (VNTR), and the primate-specific <i>Alu</i> SINE. LINEs, SINEs, ERVs, and SVAs are the major classes of retrotransposon.	45
1.4	The basic mechanism of retrotransposition leading to multiple retrotransposon copies in the genome.	46

1.5	The three main classes of retrotransposon in mammalian genomes. ERVs are characterised by long terminal repeats (LTRs), which are not found in LINEs or SINEs. ERVs and LINEs are autonomous. ERVs encode the viral <i>gag</i> and <i>pol</i> genes for retrotransposition, while LINEs have two open reading frames (ORFs) coding for the necessary proteins. SINEs are not autonomous, and often rely on LINEs for retrotransposition. Figure adapted from [36] and [37]. . . .	48
1.6	The target-primed reverse transcription (TPRT) mechanism used by LINEs for retrotransposition, adapted from [37].	50
1.7	The mechanism of retrotransposition used by ERVs, adapted from [37].	53
1.8	The addition of a methyl group to cytosine to form methylated cytosine/5-methyl cytosine (5mC). Figure from [59].	55
1.9	The DNA methyltransferase (DNMT) enzymes in mouse. Dnmt1 maintains already-established methylation. Dnmt2 and Dnmt3A/3B establish methylation de novo. Dnmt3L does not methylate directly, but is an important co-factor facilitating methylation by the catalytic methyl transferases. Figure from [60].	55

1.10	Methylation levels during mouse development. Cells undergo two waves of reprogramming. This first occurs after fertilisation. The paternal pronucleus undergoes rapid demethylation (blue), followed by passive loss of methylation in the maternal pronucleus. Methylation is then re-established in the inner cell mass (ICM). In the primordial germ cells (PGCs, shown in green), there is a second wave of reprogramming, which establishes sex-specific methylation patterns. During the periods of low methylation, retrotransposons are free from methylation-based control. Figure from [70].	57
1.11	A cartoon showing the structure of nucleosomes. DNA (red) wraps around the histone proteins (yellow). The tails can be chemically modified, which results in epigenetic regulation. Figure from [9]. . .	58
1.12	The Agouti viable yellow (Avy) example of a retrotransposon influencing gene expression. (A) The insertion of an intracisternal A particle (IAP), a mouse-specific ERV, upstream of the Avy gene leads to ectopic expression if the ERV is unmethylated. This leads to a metastable epiallele. (B) Genetically identical individuals with different epigenetic states leading to distinct phenotypes. Figure from [98].	64
1.13	A cartoon showing L1 transduction. RNA polymerase reads through the weak termination signal in the L1 until an alternative is found. The whole transcript is then retrotranscribed and inserted into a new location. Thus, a new copy of both the L1 and neighbouring sequence is produced. This can potentially create gene copies in new locations.	67

1.14	Syncytin genes in mammalian lineages. In a remarkable example of convergent evolution, each lineage has independently evolved syncytin genes from ERVs. Figure from [52].	71
1.15	The process by which new retrocopies are formed as a result of the activity of reverse transcriptase (RT) from a retrotransposon. Instead of targeting the retrotransposon RNA, the RT enzyme targets mRNA from a gene. New DNA is then synthesised using the mRNA as a template, and this is inserted into the genome at a new location. Hence, a new partial copy of the original gene arises.	74
1.16	The major steps involved in extracting RNA from a sample, as the first step in RNA-seq.	86
1.17	The major steps involved in preparing a library for RNA-seq, starting from purified RNA. Partially adapted from [180].	87
2.1	The strain/sex/cell type combinations used in the BLUEPRINT WP11 datasets.	97
3.1	(A) An example of histone modification visualisation. Each blue/red bar represents a ChIP-seq peak. Black arrowheads show the direction of transcription. (B) An example of methylation visualisation. In both (A) and (B), blue represents B cells and red T cells. Green blocks are exons.	117
3.2	The effect of median normalisation on the distribution of protein abundance values. Missing values were ignored. The normalisation process makes the different samples more comparable.	118

3.3	A comparison showing the potential false positives created by using a naive intersection method instead of an exon/block-aware method.	120
3.4	A screenshot from the UCSC Genome Browser showing the RepeatMasker tracks. A LINE element (blue) has been split by an ERV insertion (green) and a simple repeat (red).	120
3.5	An example visualisation of the retrotransposon content of a transcriptome. Agglomerative clusters are shown in the left-hand bar. Each row in the central plot represents a single transcript. Each of the main retrotransposon families has an associated colour (LINEs, green; ERVs, red; SINEs, blue), with shades representing subfamilies, shown in the colour bar at the top. White is non-retrotransposon sequence, labelled “UNIQ”.	123
3.6	Receiver operating characteristic (ROC) plots showing the performance of the cluster comparison algorithm with different covariance values and different score thresholds. Sigma represents the value used to construct the covariance matrix for the MVN distributions. As the covariance increases and clustering becomes more noisy, the false positive rate (FPR) increases; however, using a score threshold of 7.0 produces optimal true positive rates (TPRs) in every case, and low FPRs.	127

3.7	BL6 retrocopy matches in CAST. The outer ring represents the BL6 genome, while the inner ring is CAST. Lines link BL6 retrocopies to their matches in CAST, after filtering on length of match and removing interchromosomal hits. A large number show significant shifts in relative position on the chromosome. Figure created using the Circos software [238].	134
3.8	The relative positions on chromosomes for retrocopies in BL6 and their matches in CAST, after filtering for length and removing interchromosomal matches. Red dashed lines indicate the cutoff for keeping a match.	135
4.1	Summary of alignments for the BLUEPRINT BL6 samples. The left-hand figure shows proportions of reads falling into each category, while the right shows the number of reads. In some cases there is significant reduction in coverage due to removal of rRNA and unmapped reads.	142
4.2	Summary of alignments for the BLUEPRINT CAST samples. The left-hand figure shows proportions of reads falling into each category, while the right shows the number of reads. As in the BL6 samples, all samples have reduced coverage due to rRNA and unmapped reads. This suggests that the method used to identify rRNA regions in CAST was effective.	143

4.3	Distribution of alignment score (AS) and number of hits (NH) for the BL6 samples. In all samples, the majority of alignments are high scoring at around 200, and the majority of reads map uniquely, indicating that the alignments are good quality.	144
4.4	Distribution of alignment score (AS) and number of hits (NH) for the CAST samples. The distributions of AS are not as good as for the BL6 samples, with no scores above 200 for any samples. This may be due to the lower quality of the CAST reference genome compared to the BL6 reference. However, scores are still high, with the majority of alignments scoring near 200. As in BL6, the majority of the CAST alignments are unique. Overall, the results still indicate a good alignment for the CAST samples.	145
4.5	The cumulative distribution of NH values in the male and female BL6 samples. While the higher NH values do not clearly differ between male and female values, the female samples have fewer uniquely mapping reads on the Y chromosome and more with at least 2 mappings.	147
4.6	Transcriptome summaries for each BL6 sample. All samples show consistent features in the reconstructed transcriptomes.	149
4.7	Transcriptome summaries for merged BL6 transcriptomes. As expected, the transcriptome merging all samples (labelled ALL) has the highest number of features in all categories. All of the visualised features are consistent across the merges.	150

4.8	Transcriptome summaries for each CAST sample. There is more variation in the samples here compared to the BL6 samples, possibly due to the lower quality of the CAST reference genome. This is particularly true of the “transcripts per chromosome” figure (top right). However, every sample follows the same pattern across the chromosomes.	151
4.9	Transcriptome summaries for merged CAST transcriptomes. As in the per-sample CAST summaries, there is more variation than in the equivalent BL6 analysis; this is to be expected given the variation between the individual samples. A notable difference from both the BL6 merges and the individual CAST samples is in the “transcripts per gene” and “exons per transcript” figures (top left, top middle). These show a much stronger bias towards single-transcript genes and single-exon transcripts. This is not the case in the individual CAST samples, suggesting that the isoforms and exon junctions found in the individual samples are not consistent, and so cannot reliably be called in the merged transcriptomes. This may be due to less reliable mapping to the lower-quality CAST genome.	152
4.10	Breakdown of comparison between BL6 merged reconstructed transcriptomes and the Ensembl reference annotation.	153

5.1	A heatmap showing the retrotransposon content of the reconstructed transcriptomes, merged across all BL6 samples. Each row represents one reconstructed transcript, and the bars in that row represent the total proportion of retrotransposon sequence in the exons of that transcript. The colours represent the different classes of retrotransposon, according to the legend at the top of the figure. Greens represent LINES; reds represent LTRs; blues represent SINEs; and white for non-RT sequence. Different shades of each colour represent specific subclasses. The blocks on the left of the figure show the results of agglomerative clustering based on the total proportion of each type of retrotransposon. The methods used to produce these figures are described in detail in Methods. In this figure, clustering reveals relatively little structure: there is one small cluster of transcripts with high ERV1 content; a larger cluster with high RT content, but a mix of classes; and the large cluster with relatively low RT content.	158
-----	--	-----

5.2	Retrotransposon content of transcripts containing >50% RT sequence.	160
-----	---	-----

5.3	Retrotransposon content of transcripts containing >90% RT sequence.	160
-----	---	-----

5.4	The proportion of each of the major retrotransposon classes with different filters applied. All RTs: all retrotransposon (RT) elements in the genome; >0%: all RTs overlapping a reconstructed transcript; >50%: RTs overlapping a reconstructed transcript with more than 50% RT content; >90%: RTs overlapping a reconstructed transcript with more than 90% RT content. As more stringent filters are applied, LTRs become enriched compared to their proportion across the genome, while SINEs are depleted.	162
5.5	The proportion of reconstructed transcripts in each gffcompare category, for different levels of retrotransposon sequence content. Transcripts matching protein-coding reference transcripts tend to have lower retrotransposon content, as expected, and novel transcripts, which are potential lncRNAs, tend to have higher retrotransposon content.	164
5.6	The distribution of correlation coefficients between expression levels of intronic StringTie transcripts and their corresponding protein-coding transcripts.	166
5.7	The distribution of correlation coefficients between expression levels of intergenic StringTie transcripts and protein-coding transcripts within 5kb.	166
5.8	The distribution of correlation coefficients between expression levels of antisense StringTie transcripts and their corresponding protein-coding transcripts.	166
5.9	The retrotransposon content of antisense transcripts.	169

5.10	Expression correlation distributions for antisense transcripts in each retrotransposon matching category.	170
5.11	Cluster comparison between B and T cells based on retrotransposon content of transcripts with more than 50% retrotransposon sequence. The central heatmap shows the similarity score between each pair of B/T clusters; a higher score means the clusters are more similar. The clusters are also shown, in the same manner as described in Methods and shown in Figure 5.2. This confirms the existence of similar clusters of transcripts with high retrotransposon content in both B and T cells.	173
5.12	Overlap in transcribed retrotransposons between liver, B cells, and T cells. (A) All transcripts with retrotransposon content. (B) Transcripts with >50% retrotransposon content. (C) Transcripts with >90% retrotransposon content. In all three cases, there is a higher degree of shared retrotransposons between B and T than with liver, suggesting lineage-specificity as well as cell-type specificity.	175
6.1	Distribution of expressed retrocopies (inner red heatmap, inner numbers) and their parent transcripts (outer blue heatmap, outer numbers) in the mouse genome. Created using the Circos software [238].	182
6.2	The proportion of retrocopies and their parents found on each chromosome, both expressed (red) and randomly selected (grey). Error bars indicate the standard deviation of proportions from 1000 random samples, each of similar size to the number in the expressed set.	183

6.3	(A) Retrocopies expressed in B and T cells. (B) Parents of retrocopies expressed in B and T cells. In both cases, there is a high degree of overlap between the cell types.	186
6.4	(A) Expressed retrocopies in B cells, T cells, and liver. (B) Parents of expressed retrocopies in B cells, T cells, and liver. Liver shows little overlap with B and T individually, but there is significant overlap between all three.	188
6.5	The possible combinations of parent transcript strand, retrocopy strand, and transcript strand. Blue lettered boxes represent exons. The addition of ' represents the reverse complement. (i) The original transcript in the genome. (ii) The parent mRNA. (iii) The retrocopy insertion and its possible transcription. Sense with respect to (wrt) the parent produces RNA equivalent to the parent. Antisense wrt the parent produces RNA complementary to the parent.	190
6.6	Retrocopy/parent alignment scores, where 1.0 represents a perfect full-length alignment (see Methods). ALL: All retrocopies. EXPR: All expressed retrocopies. RANDOM_ALN: Negative control where retrocopies are aligned to randomly chosen parent transcripts. Expressed retrocopies are clearly biased towards high scores compared to all retrocopies.	193
6.7	An example of a retrotransposon indel (RTI) with a retrocopy transcribed from inside it. A retrocopy of the <i>Lsm5</i> gene has inserted into an ERVK element, creating an RTI with the retrocopy inside. The reconstructed transcriptomes show that this retrocopy is transcribed across all of the BL6 samples.	196

6.8	Classifications of RTIs by contents. In this case, “pseudogene” is synonymous with retrocopy. “NONE” means that the RTI could not be classified using repeats and retrocopies, and may contain other sequence. There is a significant enrichment for retrocopies in the expressed RTIs (Table 6.5).	197
6.9	Fold change (FC) of retrocopy parents with a retrocopy expressed in either B cells or T cells. Positive log FC values indicate upregulation in T cells compared to B cells. There is a bias towards upregulation in the presence of a retrocopy in each case.	201
6.10	A scatter plot showing the log fold change (FC) against the alignment score between retrocopy and parent, for retrocopy parents with a retrocopy expressed in either B cells or T cells. Spearman’s rho values do not show strong correlations in either cell type. The bias in expressed retrocopies towards high sequence similarity will skew these results, however.	202
6.11	Fold change (FC) of retrocopy parents with a retrocopy expressed in either liver or lymphocytes. Positive log FC values indicate upregulation in lymphocytes compared to liver. The small number of liver-specific parents make the results less clear. Examining the shared parents, it appears that there is a general upregulation in lymphocytes.	204
6.12	Fold change (FC) between B and T values. Immunoglobulin kappa variable transcripts are highlighted in red. Negative log FC indicates upregulation in B cells compared to T cells.	205

6.13 Distributions of normalised protein abundance levels for different sets of proteins. EXPR: Proteins with an expressed retrocopy. UN-EXPR: Proteins with a retrocopy that is not expressed. ALL: All proteins for which data is available.	210
--	-----

List of Tables

1.1	A summary of the major classes of non-coding RNAs [9–12].	40
1.2	Examples of well-studied human lncRNAs that have been functionally characterised [10, 26–28]. All of these have an orthologue in mouse, although <i>Hotair</i> may not be functional in mouse [29, 30]. <i>IGF2R</i> : insulin-like growth factor 2 receptor. <i>LSD1</i> : Lysine-specific demethylase 1. <i>PRC2</i> : Polycomb repressive complex 2. <i>HOX</i> : Homeobox.	43
1.3	A summary of the retrotransposon content of the mouse genome according to RepeatMasker [8].	47
3.1	The classes used by gffcompare for query transcripts in comparison to reference transcripts.	114
3.2	The results of the CAST rRNA discovery pipeline compared to the BL6 annotation.	115

3.3	A representative subset of the results from testing the cluster comparison algorithm (full results can be found in Online Resources). This confirms the observations from Figure 3.6 that 7.0 is a good choice of score threshold, as it produces optimal TPRs with minimal FPRs (usually zero or close to zero).	128
5.1	A summary of the BL6 reconstructed transcriptomes and the retrotransposon (RT) content of each. A very small percentage of transcripts overlapping retrotransposons contain more than 50% retrotransposon content, and about half of these contain more than 90% retrotransposon content.	157
5.2	The number of individual retrotransposon (RT) elements overlapped by transcripts in the reconstructed transcriptome across all B and T samples. SINEs dominate when including all transcripts with retrotransposon content, but not when filtering on retrotransposon content percentage. In particular, the number of ERV elements increases rapidly, which is reflected in the sequence content (Figures 5.1, 5.2, and 5.3).	161

5.3	Upper table: Anderson-Darling (AD) statistic values comparing the distributions of Spearman’s rho values shown in Figures 5.6, 5.7, and 5.8. Lower table: critical values for the AD statistic at different significance levels. For intronic and intergenic transcripts, none of the comparisons have AD statistics exceeding any of the critical values, and there is no evidence to suggest that any of their distributions are significantly different. For antisense transcripts, the distribution for the “Without RTs” category is significantly different from both other categories. These results are consistent with visual inspection of the distributions, and confirm that the presence of RTs correlates with higher Spearman’s rho values in sense/antisense pairs.	167
5.4	Anderson-Darling statistic values comparing the distributions of Spearman’s rho values shown in Figure 5.10. All of these values exceed the critical value for 1% significance, indicating that all four distributions are significantly different from each other. In particular, the distribution of values for the “RTs Match” category is very different from the distributions for the “RTs Mismatch” and “Without RTs” categories.	171
5.5	The results of a chi-squared test comparing the number of retrotransposons in each Venn category for all transcribed retrotransposons (expected), and for those in a transcript consisting of more >50% retrotransposon content (observed) (see Methods). The results are significant, and the observed versus expected values suggest that with the 50% filter the individual retrotransposons are more likely to be cell-type specific.	174

5.6	The results of a chi-squared test comparing the number of retrotransposons in each Venn category for retrotransposons in a transcript consisting of more >50% retrotransposon content (expected), and those in a transcript consisting of more >90% retrotransposon content (observed) (see Methods). The results are similar to those shown in Table 5.5, suggesting that retrotransposons are more likely to be cell-type specific with the more stringent filter.	176
6.1	Comparison of retrocopy transcripts to Ensembl reference transcripts according to gffcompare, across all BL6 samples. Codes “i”, “u” and “x” can be regarded as novel transcripts (highlighted in blue, 1,010 in total), while the remaining codes indicate a type of match. However, only 9 match exactly (code “=”).	181
6.2	The results of a gene ontology (GO) analysis of the genes on chromosome 6. There is significant enrichment for genes related to the immune response, suggesting that chromosome 6 will be in an open conformation in lymphocytes. GO analysis carried out using the online GO Enrichment Analysis tool from the Gene Ontology Consortium [249].	184
6.3	Number of retrocopy transcripts, expressed retrocopies, and corresponding parent transcripts across all BLUEPRINT samples, in B cells, in T cells, and in liver.	187

6.4	The number of expressed retrocopies across all BLUEPRINT BL6 samples falling into each strand combination. A chi-squared test shows that there is a very clear enrichment in the categories leading to reRNA complementary to the parent (highlighted in blue). . . .	191
6.5	The results of a chi-squared test comparing the contents of all RTIs to those that are expressed. To obtain the expected values, the proportions falling into each category across all RTIs were multiplied by the total number of expressed RTIs. This shows a clear and significant enrichment for retrocopies in the expressed RTIs. . . .	198
6.6	The results of a chi-squared contingency test comparing expression of retrocopies in BL6 and their location inside an RTI. The observed values differ significantly from the expected values, suggesting that retrocopy expression and location inside an RTI are not independent.	199
6.7	Parent transcripts with at least one retrocopy expressed cell type-specifically, and with upregulation in the presence of retrocopy expression.	206
6.8	The results of Anderson-Darling (A-D) tests comparing protein abundance distributions (Figure 6.13). The "critical value" represents the value of the A-D statistic with a 1% significance level.	209

6.9	The results of a chi-squared contingency test comparing expression of retrocopies in BL6 and their conservation in CAST. The observed values differ significantly from the expected values, suggesting that fewer BL6 expressed retrocopies are conserved than would be expected by chance. However, the significance level of this test is not very low, and the difference between the observed and expected values is not large, so this is not a strong result.	212
6.10	The results of a chi-squared contingency test comparing parents of expressed retrocopies in BL6 and the parents of retrocopies conserved in CAST. The observed values differ significantly from the expected values, suggesting that there are more parents with both an expressed retrocopy in BL6 and a conserved retrocopy in CAST.	213
6.11	The results of a chi-squared contingency test comparing parents of expressed retrocopies in BL6 and the parents of retrocopies conserved and expressed in CAST. The observed values differ significantly from the expected values, suggesting that there are more parents with both an expressed retrocopy in BL6 and a conserved and expressed retrocopy in CAST.	214

6.12	The results of a chi-squared contingency test comparing cell-shared parents of expressed retrocopies in BL6 and the parents of retrocopies conserved and expressed in CAST. That is, a retrocopy parent which has at least one retrocopy expressed in all three BL6 cell types will fall into the “BL6 Expressed and Shared” category; a parent with a retrocopy expressed in only one or two cell types will fall into the “BL6 Expressed, not Shared” category. The observed values differ significantly from the expected values, suggesting that there are more parents with both an expressed retrocopy in BL6 across B, T, and liver cells, and a conserved and expressed retrocopy in CAST.	216
------	--	-----

Chapter 1

Introduction

The 1950s and 60s were a remarkable time for molecular biology. Arguably the most famous discovery of this period was the structure of DNA by Crick and Watson in 1953, based on work by Wilkins and Franklin [1]. In 1961, Jacob and Monod proposed the Jacob-Monod model, now a textbook example of gene regulation, and laid out a theory for the role of messenger RNA (mRNA) [2]. In 1961, two studies demonstrated the role of mRNA in protein synthesis [3,4], partly based on previous work from several authors [5]. Together, these findings laid the foundation for the “central dogma” of molecular biology: DNA is transcribed into RNA, RNA is translated into protein. This is an appealing model in its simplicity, and serves as a good starting point for the student of molecular biology. Unfortunately, it is insufficient for explaining many other phenomena in genetics and cell biology. In particular, the related genetic phenomena of non-coding RNA (ncRNA) and retrotransposons require fundamental revisions to this model.

The central dogma, as stated above, suggests that the role of RNA is only to transmit genetic information, as an intermediary between DNA and protein. In

truth, there are many kinds of RNA that do not code for a protein, known as non-coding RNA, which instead fulfill a role as RNA. The advent of high-throughput sequencing (HTS) of RNA led to the identification of thousands of distinct ncRNAs. A few examples have well-characterised functions, in the classical sense: they have an observable knockout phenotype, or a known biochemical mechanism in the cell. These examples demonstrate the potential of ncRNA, but are only a tiny minority, leaving tens of thousands of positively identified transcripts without a clear role, if any. This leads to the question of whether every single ncRNA is functional. Indeed, the answer may call for a revised idea of molecular function. Is it possible that every individual ncRNA has an observable knockout phenotype or distinct molecular pathway? Some authors have suggested alternatives, whereby ncRNAs act in concert; hence, removal of a single ncRNA would have little impact, but the removal of many would be severely deleterious.

The central dogma, as stated above, only goes in one direction, from DNA to RNA to protein; it does not allow for the passage of information back from RNA to DNA. In 1950, Barbara McClintock published a paper describing DNA segments in maize that could move within the genome [6]. While fascinating, the importance of these "jumping genes" was unclear. McClintock herself speculated that these elements controlled gene expression, although this idea was not immediately accepted [7]. These jumping genes are now known as transposable elements (TEs): DNA sequences that can move within the genome, or can copy themselves to a new location. As with ncRNAs, HTS revealed the scale of their presence: TEs have been found in nearly all species studied, and can account for significant proportions of the genome. In particular, mammalian genomes have a high proportion of retrotransposons, TEs that produce new copies of themselves

in the genome through retrotranscription, a process that uses RNA as a template to create new DNA. For example, about 40% of the human genome is made up of TEs, nearly all of which are retrotransposons [8]. It is also widely accepted that they are vital components of the genome, as McClintock correctly predicted, and as shown by a few well-studied examples. However, as with ncRNAs, it is not yet fully understood how many are functional, or even in what sense they might be functional. In addition, their presence in the genome has led to a number of other effects, such as gene duplication, altered regulatory networks, and the emergence of novel genes.

Non-coding RNA and retrotransposons represent two genome-scale phenomena, not yet fully understood, but undoubtedly important. Both have been linked to human disease, and cancer in particular. It is particularly fascinating to note that many studies have now established a link between ncRNAs and retrotransposons, and so the study of one almost inevitably leads to the study of the other. At the level of basic science, they are certainly worth studying: huge numbers of mysterious transcripts, large proportions of genomes, across almost all species studied. Beyond the motivation of scientific curiosity, their relevance in medical science is now beyond question.

1.1 Long Non-Coding RNA

As discussed above, thousands of distinct non-coding RNAs have been discovered across many species. They can be classified into several different kinds, based on size and function (see Table 1.1).

Most relevant to this work are the long non-coding RNAs, a large class of

Name	Abbr.	Size (nt)	Function
Ribosomal RNA	rRNA	120 - 5,025	Ribosome components; protein synthesis
Long non-coding RNA	lncRNA	>200	Various regulatory roles; see below
Transfer RNA	tRNA	70 - 90	Translation of codons to amino acids
Piwi-interacting RNA	piRNA	23 - 31	suppression
Micro RNA	miRNA	22	Post-transcriptional gene regulation via base-pairing
Small interfering RNA	siRNA	22	Post-transcriptional gene regulation via RNA interference pathway
Small nuclear RNA	snRNA	100 - 300	Pre-messenger RNA processing
Small nucleolar RNA	snoRNA	70	Guiding RNA nucleotide modifications

Table 1.1: A summary of the major classes of non-coding RNAs [9–12].

ncRNAs usually defined as being more than 200bp in length [11]. Several thousand distinct lncRNAs have been identified in each of the mouse and human genomes (as well as other species), and they generally have the following features:

1. Highly tissue-specific expression patterns [11, 13, 14]
2. Low expression compared to coding RNAs [14, 15]
3. Low sequence conservation across species [14, 16], although, recent work by Hon and colleagues suggests higher levels of conservation than previously thought [17]

The advent of high-throughput transcriptomics radically changed our understanding of lncRNAs, revealing that thousands of distinct lncRNAs exist. However, low expression levels and tissue specificity make them particularly difficult to detect using sequencing approaches, notwithstanding the usual biological noise that affects studies of this kind (e.g., cell cycling, mixed cell populations). Differences in experimental design, such as sequencing depth and choice of analysis tools, can also alter the transcripts detected. It is therefore hardly surprising that estimates of the number of lncRNAs vary widely between studies (Figure 1.1) [18], and there are still no consensus lncRNA annotations in mouse or human.

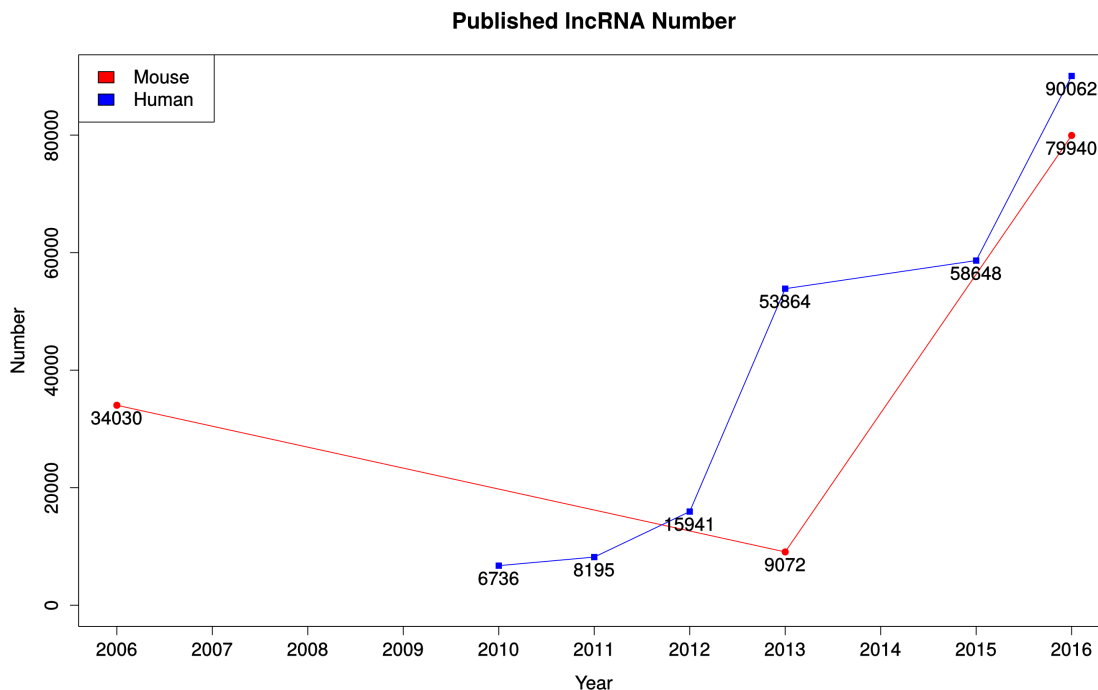


Figure 1.1: The number of lncRNAs found in human and mouse in different transcriptomics studies. Figure based on Table 1 in [18], using data from [13–15,19–23]

While annotations continue to expand and improve the catalogue of known lncRNAs, the question of function is still unanswered. The sheer number of lncR-

NAs present in the mouse and human genomes makes classical genetic approaches difficult, although recent studies have demonstrated techniques that can assess the function of multiple lncRNAs [17, 24, 25].

A minority of well-studied lncRNAs have been shown to perform a role in the cell (such as those shown in Table 1.2), and these illustrate the regulatory potential of lncRNAs. These are not the only examples, and it is likely that more will be discovered as lncRNA study continues. However, these still account for a small proportion of the lncRNA transcripts identified in model organisms. Some studies have suggested that most lncRNAs do not have specific roles as individuals, but instead work together. For example, Rinn *et al.* have suggested that lncRNAs are used as anchors to alter chromatin structure [18]. Such theories suggest that the majority of lncRNAs do not have distinct, individual functions, but instead act together in concert. Traditionally, to say a transcript is functional would imply that its absence is associated with an observable phenotype. It may be that to understand lncRNAs requires a different perspective, in which individual transcripts are not in themselves functional, and to remove them would have little or no effect; but to remove all such transcripts would have severe consequences.

For the time being, it is unknown whether every single one of these transcripts has a distinct function; whether lncRNAs work together as a whole; or whether they are transcriptional noise. It may be a mixture of all three possibilities. While debate and study of their function continues, there is a common finding across recent surveys of lncRNAs: they have a strong link with transposable elements, particularly retrotransposons.

Name	Description	Chromosome	Function
<i>XIST</i>	X inactive specific transcript	X	Coats and silences one copy of the X chromosome during X chromosome inactivation
<i>AIRN</i>	Antisense <i>IGF2R</i> RNA	6	Induces imprinting of gene cluster including <i>IGF2R</i>
<i>HOTAIR</i>	HOX transcript antisense intergenic RNA	12	Binds <i>LSD1</i> and <i>PRC2</i> to repress <i>HOXD</i> gene cluster in <i>trans</i>
<i>FIRRE</i>	Functional intergenic repeating RNA element	X	Maintains repressive chromatin
<i>MEG3</i>	Maternally expressed gene 3	14	Possible tumour suppressor

Table 1.2: Examples of well-studied human lncRNAs that have been functionally characterised [10, 26–28]. All of these have an orthologue in mouse, although *Hotair* may not be functional in mouse [29, 30]. *IGF2R*: insulin-like growth factor 2 receptor. *LSD1*: Lysine-specific demethylase 1. *PRC2*: Polycomb repressive complex 2. *HOX*: Homeobox.

1.2 Retrotransposons in the Mouse Genome

1.2.1 Introduction to the Repetitive Genome

Repetitive DNA is defined as a nucleotide sequence that appears multiple times in the genome. These sequences are often simply termed “repeats”, and can be broadly divided into four classes [8] (Figure 1.2):

- Simple repeats: a short nucleotide sequence (e.g., T, CGG) repeated several times
- Tandem repeats: a longer (100 - 200 base pair (bp)) sequence duplicated

several times at the same location

- Segmental duplications: large blocks (10 - 300 kilobases) that have been copied to another region of the genome
- Interspersed repeats: copies of sequences of varying length (100 - 10,000 bp) at several locations throughout the genome

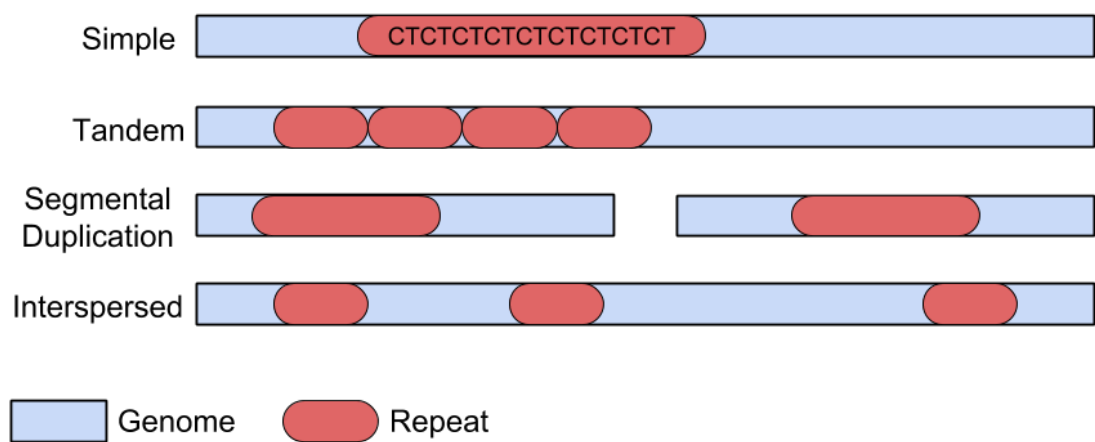


Figure 1.2: The four broad classes of repeats.

The majority of interspersed repeats are transposable elements (TEs): DNA elements characterised by an ability to move around the genome, and hence give rise to interspersed repeats. TEs can be either DNA transposons, which move via a “cut and paste” mechanism, or retrotransposons, which move via a “copy and paste” mechanism using an RNA intermediate that is retrotranscribed [31].

Approximately 45% of the mouse genome has been identified as repetitive by RepeatMasker (Figure 1.3) (see Datasets for a discussion of the RepeatMasker software). The vast majority of repetitive elements are retrotransposons, with only a small percentage of the genome identified as simple repeats, tandem repeats,

or DNA transposons. A similar distribution is observed in the human genome, although with a higher total percentage of repetitive DNA (52.5%, Figure 1.3). While the distributions are similar and the broad classes of repeat are shared between them, the exact subclasses are often species- or lineage-specific (such as the primate-specific *Alu* retrotransposon).

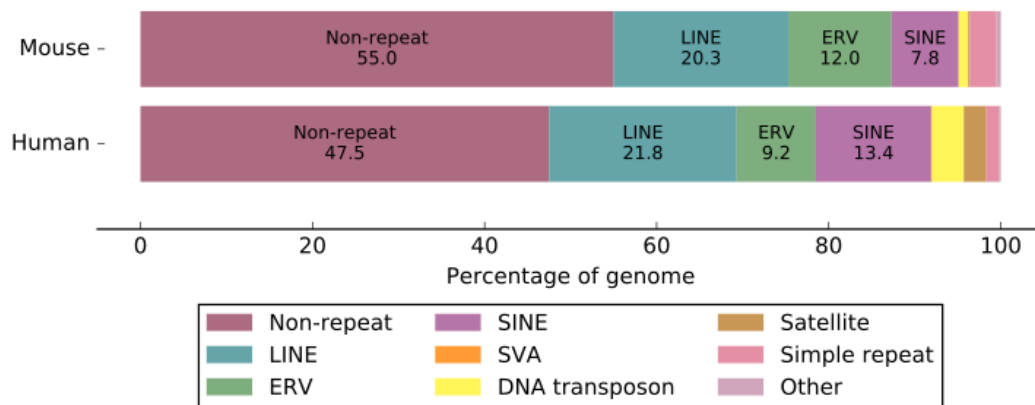


Figure 1.3: The repetitive DNA content of the mm10 mouse and hg38 human reference genomes, as identified by RepeatMasker [8]. LINE: long interspersed element. ERV: endogeneous retrovirus. SINE: short interspersed element. SVA: SINE VNTR *Alu*; a primate-specific retrotransposon composed of a SINE, a variable number tandem repeat (VNTR), and the primate-specific *Alu* SINE. LINES, SINEs, ERVs, and SVAs are the major classes of retrotransposon.

1.2.1.1 Retrotransposons

Retrotransposons (RTs) are characterised by their “copy and paste” mechanism, which has resulted in many hundreds or thousands of RT copies scattered throughout their host genomes. While each type of retrotransposon has a different exact mechanism for retrotransposition, the steps are broadly similar [31] (Figure 1.4):

1. The RT is transcribed into RNA

2. This RNA is reverse transcribed back into DNA by a reverse transcriptase enzyme
3. The newly synthesised DNA is inserted into the host genome at a new location

Thus, a new copy of the retrotransposon has been created. The use of reverse transcriptase to synthesise DNA from an RNA template is the key feature that differentiates RTs from other types of transposable element.

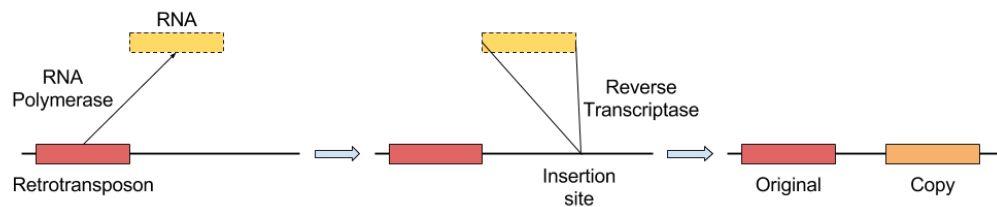


Figure 1.4: The basic mechanism of retrotransposition leading to multiple retrotransposon copies in the genome.

In practice, this process is often imperfect, causing truncations and other mutations in the new retrotransposon copy [32,33]. RTs are also subject to the usual mutations that affect the host genome as a whole. This means that many of the RTs identified in the genome are partial or otherwise imperfect copies that are no longer competent for retrotransposition [34,35]. Over evolutionary time, new RTs have emerged and older ones have decayed, leading to species-specific retrotransposons [31]. Examination of shared retrotransposons between species, along with the degree to which their copies are decayed, can be used to estimate the age of a retrotransposon type.

Retrotransposons are described as either autonomous or non-autonomous. Autonomous retrotransposons encode all of the cellular machinery needed to copy

themselves (except for RNA polymerase), while non-autonomous RTs rely on machinery from another source, such as an autonomous RT. It should be noted that classification as autonomous does not guarantee that a given RT element is able to successfully retrotranspose by itself - as discussed above, acquired mutations have rendered the majority of copies incompetent for retrotransposition. Rather, it means that the complete and unaltered version of this element would be able to retrotranspose using its own cellular machinery.

Retrotransposons (RTs) in mammals can be broadly divided into three types (Figure 1.5), and their presence in the mouse genome is summarised in Table 1.3:

- Long INterspersed Elements (LINEs)
- Short INterspersed Elements (SINEs)
- Endogenous Retroviruses (ERVs)

Each type can be recognised by its structure and mechanism of retrotransposition, and each has originated from a different source.

	Autonomous?	Length (bp)	% of mouse genome	Number of elements in mouse genome
LINE	Yes	500 - 8,000	20.3	713,890
SINE	No	100 - 300	7.8	1,451,137
ERV	Yes	200 - 5,000	12.0	738,286

Table 1.3: A summary of the retrotransposon content of the mouse genome according to RepeatMasker [8].

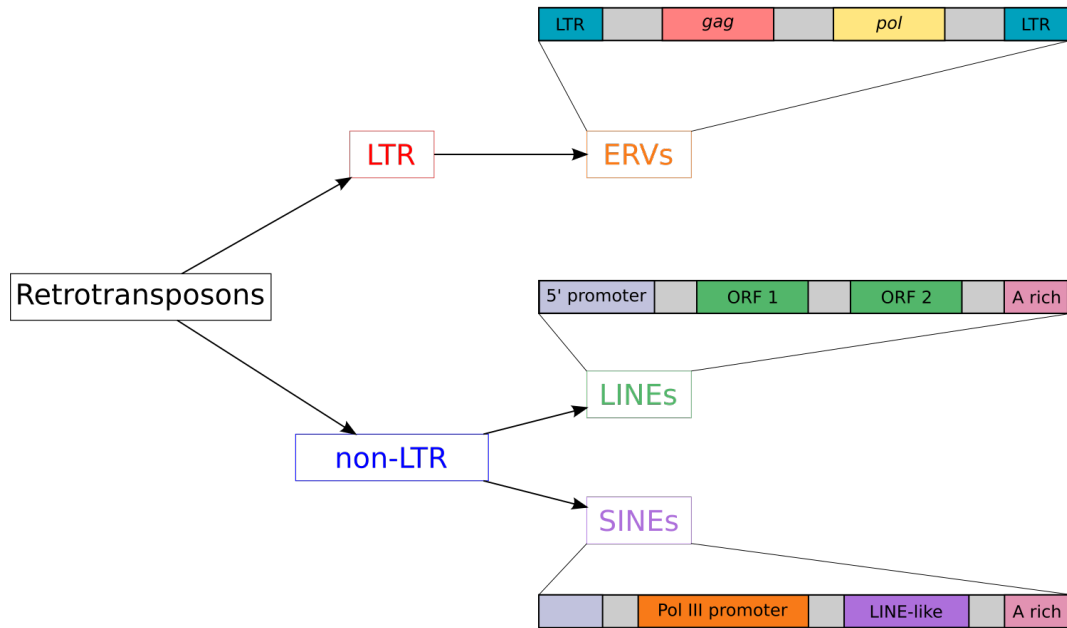


Figure 1.5: The three main classes of retrotransposon in mammalian genomes. ERVs are characterised by long terminal repeats (LTRs), which are not found in LINEs or SINEs. ERVs and LINEs are autonomous. ERVs encode the viral *gag* and *pol* genes for retrotransposition, while LINEs have two open reading frames (ORFs) coding for the necessary proteins. SINEs are not autonomous, and often rely on LINEs for retrotransposition. Figure adapted from [36] and [37].

1.2.1.2 LINEs

LINEs are autonomous, and have been highly successful in eukaryotic genomes, particularly in mammals [8] [38]. In mammals, LINEs can be classified as LINE-1 (L1) or LINE-2 (L2), although L1s dominate, and the few remaining L2s are fossils that are not competent for retrotransposition.

Complete LINEs typically have the following structure (Figure 1.5) [39]:

- A 5' untranslated region (UTR) that contains an internal RNA Polymerase II (RNAPII) promoter

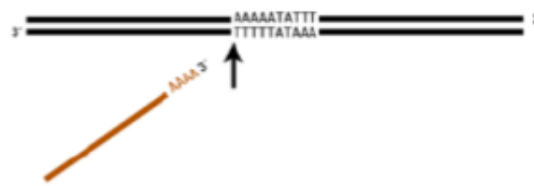
- Open reading frame (ORF) 1, which encodes an RNA binding protein
- ORF2, which encodes a protein with endonuclease and reverse-transcriptase functions
- A 3' adenosine-rich region

The proteins encoded by ORF1 and ORF2 are both crucial for retrotransposition. Both exhibit a strong *cis*-preference and so tend to associate with the RNA molecule that encoded them, in an attempt to ensure that the LINE itself is retrotransposed [40–42].

LINEs retrotranspose through the target-primed reverse transcription (TPRT) mechanism [43] (Figure 1.6). This mechanism typically produces target-site duplications, 5' truncation of the LINE, and A-rich sequence at the 3' end; it can also lead to inversions of the L1 [31–33, 43]. This propensity for errors has left only a few active copies remaining (about 100 in humans) [34].

The origin of LINEs is still unclear. There is evidence that they may have evolved from group II introns, mobile genetic elements found in bacterial and mitochondrial genomes (reviewed in [44]). Group II introns and LINEs have similar reverse transcriptases, and both use the TPRT mechanism for retrotransposition. However, it is difficult to distinguish between this case and the possibility that LINEs and group II introns share a common ancestor.

1a. RNA intermediate (brown) associates with an A/T-rich region
1b. Nuclease makes single-stranded break next to a run of thymines (arrow)



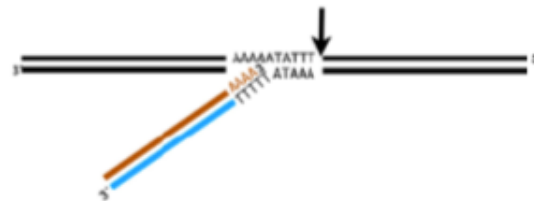
2. Poly(A) at 3' end of RNA base pairs with DNA at the break



3. Reverse transcriptase synthesises DNA (blue) using RNA template



4. Nuclease makes another break on the opposite strand, a few nucleotides away



5. Template RNA removed by RNase H



6. Synthesis of opposite strand begins



7. Host enzymes complete integration; process results in target site duplications (TSDs) either side of the insertion



Figure 1.6: The target-primed reverse transcription (TPRT) mechanism used by LINEs for retrotransposition, adapted from [37].

1.2.1.3 SINEs

SINEs are not autonomous, and instead depend on LINE machinery for retrotransposition [45, 46]. Some SINEs have sequence that closely matches that of a particular LINE, increasing the likelihood that the LINE proteins will retrotranspose the SINE instead [47]. As such, their mechanism of retrotransposition is essentially the same as for LINEs, but with SINE RNA instead of the original LINE. Despite (or perhaps because of) their reliance on LINEs, SINEs have been highly successful in eukaryotic genomes.

SINE structure is as follows [48] (Figure 1.5):

- An RNA Polymerase III (Pol III) promoter that initiates transcription
- LINE-like sequence used for retrotransposition, although this is not always present
- An A-rich 3' tail

However, there can be significant variation in structure between individual SINEs. As the name suggests, they are shorter than other retrotransposons, usually between 80 and 500 bp [48].

SINEs are thought to be derived from cellular RNA combined with other sequence, such as partial LINE elements. In mouse, there are four main subclasses of SINE: B1, B2, B4, and ID, in order of abundance. These are derived from tRNA (ID, B4, B2) [49, 50] or 7SL RNA (B1, B4) [50].

1.2.1.4 ERVs

ERV retrotransposons are characterised by the presence of a long terminal repeat (LTR) at each end of the retrotransposon [51,52]. Full length ERVs, partial ERVs, and solo LTR elements have all been identified in the genome [8,52]. The full ERV structure is as follows (Figure 1.5):

- The 5' LTR, which initiates transcription
- *Gag*, *pol*, and *env* genes that encode viral proteins for retrotransposition
- The 3' LTR, which terminates transcription

As the 5' and 3' LTRs are identical, they can undergo homologous recombination, excising the viral genes in between and leading to solo LTRs [53]. The mechanism of retrotransposition for LTR elements is summarised in Figure 1.7.

As the name suggests, ERVs are thought to have originated from retroviruses that infected the ancestral germline and inserted a copy of their viral genome [54]. These copies lost the ability to transfer horizontally from cell to cell, instead becoming restricted to transfer within the genome of a single cell. New ERVs have emerged within lineages, giving rise to multiple related families of ERVs within any given species. In the mouse, for example, there are three main classes of ERVs, which can be further subdivided into 13 subclasses [55]. While the behaviour of each class is broadly similar, their age and activity levels vary.

1. RNA intermediate (brown) has tRNA (blue) bound to the primer binding site (PBS) near the 5' end



2. Reverse transcriptase begins DNA synthesis at the PBS, starting with the unique sequence (U5, red) and a repeat sequence (R, green)



3. The new copy of R base pairs with complementary sequence at the 3' end



4. DNA synthesis is completed



5. Template RNA is removed, except for a fragment to prime synthesis of second strand



6. Second strand synthesis for unique 3' sequence (U3, purple), R, and U5



7. Reverse transcriptase switches template using complementarity of R and U5



8. Bidirectional synthesis creates complete ERV with LTRs at both ends



9. Integrase inserts new ERV into host genome. Transcription initiated at one LTR and terminated at the other creates template RNA with R at both ends



Figure 1.7: The mechanism of retrotransposition used by ERVs, adapted from [37].

1.2.2 Silencing of Retrotransposons

Active retrotransposons pose a great threat to genome stability, potentially causing severe deleterious mutations. Germline mutations could result in inviable offspring, and retrotransposon activity has been linked to more than 90 genetic diseases in humans [56]. L1 activity in somatic cells has been linked to numerous types of cancer, although it is unclear whether this is a cause of cancer or an effect of a more fundamental loss of regulation [57]. In order to limit the damage done by retrotransposons, mammals (and many other species), have evolved an array of mechanisms to repress retrotransposons. As an organism develops, different mechanisms become important, acting in complementary fashion to ensure continuous silencing of retrotransposons. The three major mechanisms found in mammals are DNA methylation, histone modifications, and Piwi-interacting RNA.

1.2.2.1 DNA Methylation

DNA methylation is one of the most important and best-studied epigenetic marks in mammals, and plays an essential role in retrotransposon suppression. Indeed, it has been suggested that DNA methylation originally evolved as a retrotransposon defence mechanism, before gaining other functions such as gene regulation and imprinting [58]. DNA methylation is the addition of a methyl group to a cytosine (C) to form 5-methylcytosine (5mC) (Figure 1.8). In mammals, this usually occurs at a cytosine-guanine dinucleotide (CpG).

In nearly all differentiated cells, retrotransposons are heavily methylated [60], and numerous studies have shown that demethylation is associated with increased retrotransposon activity, in a variety of contexts. In mouse, failure to establish or

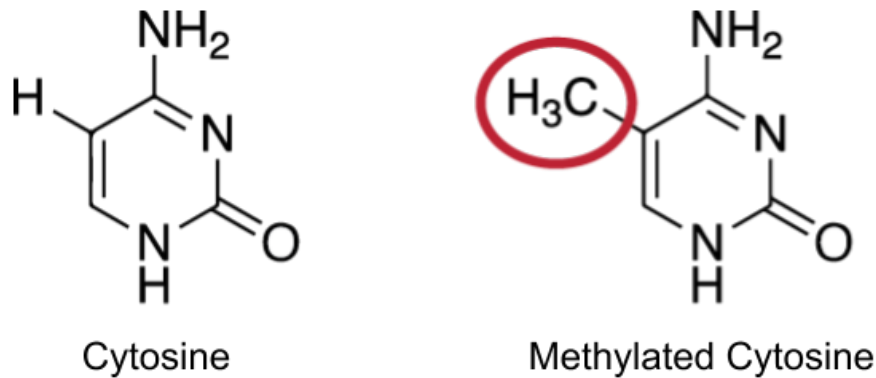


Figure 1.8: The addition of a methyl group to cytosine to form methylated cytosine/5-methyl cytosine (5mC). Figure from [59].

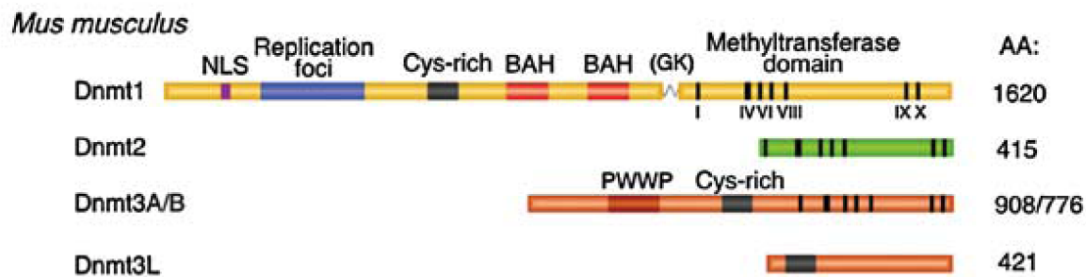


Figure 1.9: The DNA methyltransferase (DNMT) enzymes in mouse. Dnmt1 maintains already-established methylation. Dnmt2 and Dnmt3A/3B establish methylation de novo. Dnmt3L does not methylate directly, but is an important co-factor facilitating methylation by the catalytic methyl transferases. Figure from [60].

maintain methylation during development results in increased L1 and ERV activity [61,62], and removal of methylation in post-natal mice leads to increased ERV activity [63]. In human cell lines, lower methylation levels seem to correlate with increased human ERVK (HERVK) activity [64]. These findings are corroborated by Reiss and colleagues, who found that LTRs with placenta-specific promoter activity are demethylated in placenta, but methylated in blood cells [65]. Studies in cancer cell lines have found similar upregulation of L1 and ERV elements in the

absence of methylation [66–68]. These studies represent a convincing and growing body of evidence that methylation is vital in controlling retrotransposons, particularly L1s and ERVs. However, removal of methylation is not associated with increased activity across all retrotransposons. For example, SINE expression in human cell lines is not affected by demethylation [69]. Lavie *et al.* examined specific HERVK elements in different cell lines and found that they were suppressed by methylation in Tera-1 cells, but T47D cells did not show increased HERVK expression, despite reduced methylation at the same elements [64]. It may be that these elements are silenced by other mechanisms, such as histone modifications (see below). Alternatively, they may have become degraded to such an extent that methylation can no longer be targeted - at which point it is also unnecessary, as the retrotransposon will likely be incapable of retrotransposition.

1.2.2.2 Histone Modifications

During early mammalian development, there is a wave of epigenetic reprogramming, when nearly all methylation is removed from the genome and re-established (Figure 1.10) [70]. This could lead to increased levels of retrotransposon activity at a time when genome stability is particularly important. However, nearly all retrotransposons remain inactive during reprogramming, as multiple mechanisms act in a complementary fashion. One of these mechanisms is histone modifications.

Histone modification is another well-studied epigenetic mechanism. In eukaryotes, DNA is packaged as nucleosomes. Each nucleosome consists of a segment of DNA wrapped around a protein complex consisting of eight histone proteins (Figure 1.11). Chemical modifications, such as methylation and acetylation, can be made to the N-terminal tail of each histone protein. These modifications then

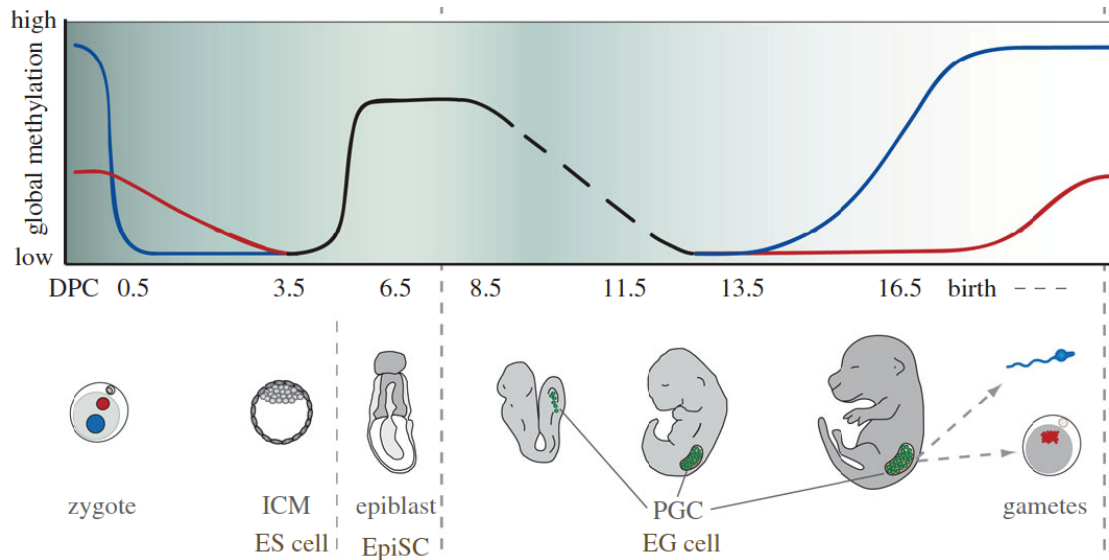


Figure 1.10: Methylation levels during mouse development. Cells undergo two waves of reprogramming. This first occurs after fertilisation. The paternal pronucleus undergoes rapid demethylation (blue), followed by passive loss of methylation in the maternal pronucleus. Methylation is then re-established in the inner cell mass (ICM). In the primordial germ cells (PGCs, shown in green), there is a second wave of reprogramming, which establishes sex-specific methylation patterns. During the periods of low methylation, retrotransposons are free from methylation-based control. Figure from [70].

affect expression of nearby regions by altering chromatin structure and/or recruiting interactors. Different combinations of histone protein, modified residue, and chemical modification lead to a wide range of effects.

The regulation of transposable elements by histone modifications is not yet completely understood, and involves several complex interactions between various DNA-binding complexes, recruiters, and histone modifiers. However, there are common findings from several groups providing some insight into the relationship between histone modifications and retrotransposon control. Two histone modifications in particular seem to be associated with retrotransposon silencing: histone 3 lysine 9 trimethylation (H3K9me3) and histone 4 lysine 20 trimethylation (H4K20me3).

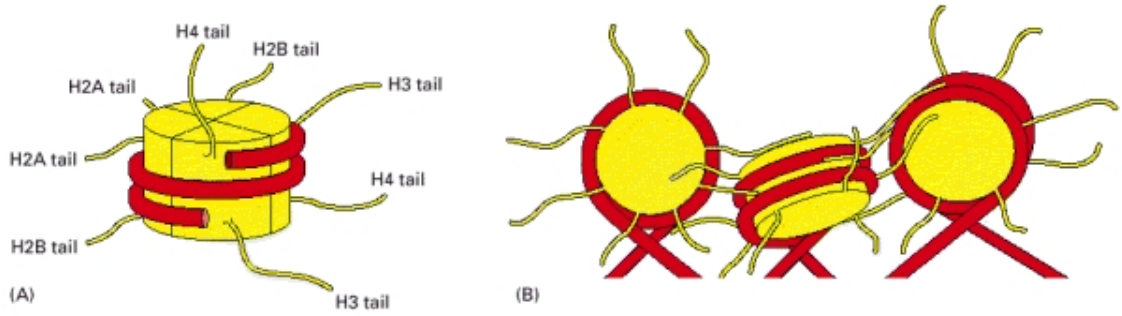


Figure 1.11: A cartoon showing the structure of nucleosomes. DNA (red) wraps around the histone proteins (yellow). The tails can be chemically modified, which results in epigenetic regulation. Figure from [9].

tion (H4K20me3). Genome-wide studies have found enrichment of one or both of these marks at ERVs in mouse embryonic stem cells (mESCs) [71–74]. Both of these marks are associated with repression, and mutants lacking enzymes that establish H3K9me3 show increased levels of retrotransposon expression, particularly ERVs [74]. This is in contrast to H4K20me3, which appears to be dispensable, with no deregulation as long as the H3K9me3 marks remain intact [73].

1.2.2.3 Piwi-interacting RNA in the Germline

As noted above, the germline undergoes epigenetic reprogramming during development. During these periods, and in general, it is important to silence retrotransposons in the germline, not least because germline insertions can become fixed in the genome. To achieve constant repression of retrotransposons in the germline, a germline-specific small RNA-mediated mechanism has evolved, based on Piwi-interacting RNAs (piRNAs). PiRNAs are a class of short non-coding RNA, between 23 and 31 bp long [12], which control transposable elements (TEs) at both the transcriptional and post-transcriptional levels. Two features of the

piRNA pathway make it extremely effective in its control of TEs: the ability to adapt to and remember new TEs, and a mechanism for amplifying piRNAs to specifically target active TEs.

As the name suggests, the piRNA silencing mechanism relies on interaction with proteins in the Piwi family, a metazoan-specific clade of the Argonaute protein family [12]. While the piRNA pathway is found throughout eukaryotes and much work has been done on piRNA in *Drosophila melanogaster* and *Caenorhabditis elegans*, this discussion will focus on piRNA in mice. The mouse genome contains three proteins in the Piwi family: MIWI, MILI, and MIWI2, which are expressed at different stages in development and correspond to different subsets of piRNA [75–78]. It should be noted that in mouse there is an unusual asymmetry between the piRNA activity in male and female germlines. In *Drosophila*, zebrafish, and several mammalian species (including humans), piRNAs are active and essential in both spermatogenesis and oogenesis [79, 80]. In mice, the Piwi proteins and piRNAs are essential for spermatogenesis and male fertility, but do not affect female fertility [75–77]. Oocytes do express the Piwi proteins [81, 82], but have very low piRNA levels [83, 84]. It is beyond the scope of this introduction to explore this interesting discrepancy, and so the rest of this section will focus on piRNAs in the male germline.

piRNAs are a highly complex group of molecules, with millions of distinct piRNAs found in mouse. However, these millions of individual sequences all map to a few regions on the genome, known as piRNA clusters [85–89]. These regions can be very large (up to 200kb), and in mouse are usually found in euchromatic domains. In mouse there are two kinds of piRNA cluster: one class which is transcribed in embryonic development, and another which is transcribed in the spermatogenic

cells of adolescent mice during meiosis (known as pachytene clusters) [81]. The former class defends against retrotransposon activity, but the function of pachytene clusters is not yet completely understood. Unless stated otherwise, the rest of this discussion will focus on the former class.

The primary role of piRNAs in the germline is to defend the genome from potentially harmful mutations caused by transposable elements. In mouse, there is evidence for both transcriptional and post-transcriptional silencing. At the transcriptional level, piRNAs and Piwi proteins are essential in establishing DNA methylation during embryonic germ cell development [81,90,91]. Failure to establish DNA methylation during this period leads to retrotransposon activation and subsequent sterility [75].

In order to understand the post-transcriptional activity of piRNAs, one must first understand their biogenesis. piRNAs can be classified as either primary or secondary, based on how they are generated. Primary piRNAs arise from precursors transcribed from piRNA clusters. These precursors are single-stranded RNA molecules that are significantly longer than piRNAs. The exact mechanism that gives rise to mature piRNAs from these precursors is not yet completely understood [12,92]. It is thought that the precursors are first cleaved into shorter piRNA precursors, which are then loaded into a Piwi protein. After loading, the piRNA precursor is trimmed and stabilised. Disruption of any of these steps leads to a reduction in the number of piRNAs, increased retrotransposon activity, and sterility [12].

piRNA clusters are enriched for retrotransposon sequence. Therefore, piRNAs contain short pieces of retrotransposon sequence, which are used to target RNA molecules transcribed from retrotransposons. These can then be cleaved by the

Piwi protein into which the piRNA is loaded. The use of retrotransposon sequence to target active retrotransposons gives piRNA a “memory” that has been likened to an adaptive immune system. If an active retrotransposon jumps into a piRNA cluster, piRNAs to target it can be generated, and so the retrotransposon can be suppressed [12, 92].

Secondary piRNAs are generated by an amplification pathway known as the ping-pong cycle. This pathway is significantly better understood than primary piRNA biogenesis, particularly in *Drosophila*. In the ping-pong cycle, TE transcripts are cleaved by Piwi proteins loaded with piRNA. These cleaved transcripts are then used by other Argonaute proteins to select transcripts from the piRNA cluster that match the cleaved transposon. These piRNA precursors are then processed to form mature piRNA that can target the active transposon, and so the cycle continues. In this way, piRNAs that target active transposable elements are amplified. The combination of memory and amplification ensures effective and adaptive suppression of any transposable element that escapes transcriptional silencing [12, 92].

1.2.3 The Role of Retrotransposons in Genome Evolution and Function

Despite the efforts to suppress their activity, the presence of retrotransposons in mammalian (and other) genomes has had far-reaching consequences for the evolution of their host organisms at a molecular level. Retrotransposons are now known to make significant contributions to the transcriptome. Alongside this, they have had a great impact on expression regulation, transcript diversity, and

epigenetic mechanisms [93].

1.2.3.1 Retrotransposons Create Regulatory Elements

Retrotransposons are a rich source of regulatory elements, including promoters, enhancers, and transcription factor binding sites (TFBSs) [52]. As retrotransposons have spread, their regulatory regions have been adopted by host genomes and used to engineer transcriptional activity, including the creation of new regulatory networks, the modification of existing ones, and the regulation of new transcriptional units. This can occur when a new RT copy inserts in or near an existing gene such that the RT regulatory elements can influence the gene's expression pattern. While this may be deleterious, in some cases it has led to beneficial effects for the host genome, and the new regulatory network has been retained.

There is overwhelming evidence for this, to the extent that there is now a catalogue of genes affected by transposable elements [94]. This catalogue lists 124 experimentally validated cases of genes influenced by transposable element-derived regulatory elements in humans, and 48 examples in mouse [94]. Summary statistics show that SINEs are responsible for 112 cases (about 50%), with LINEs and LTRs accounting for the remainder, except for the 7 ($\sim 3\%$) due to DNA transposons. About 75% of the regulatory effects are due to promoters (primary or alternative), alternative splicing, and alternative polyadenylation signals. When broken down by transposable element type, LTRs are primarily responsible for new promoters, while SINEs cause alternative splicing. Alternative splice sites and promoters account for 50% and 20% of LINE effects. In addition to the validated examples, genome wide studies have identified many thousands of possible sites where transposable elements influence genes [94–96]. Below, I describe some examples that

illustrate how retrotransposon regulatory regions can influence transcription.

The LTR regions of ERVs contain promoters used to initiate their own transcription as the first step in retrotransposition. If an ERV inserts near an existing gene, the LTR promoter can act as an alternative promoter for that gene, or can become the sole promoter. An example of the former is found in the human p63 gene. This gene was known to eliminate oocytes that had suffered from DNA damage, but an equivalent activity in the male germline had not been identified [97]. Beyer *et al.* identified a novel p63 transcript expressed specifically in testis, with a transcriptional start site (TSS) inside an upstream LTR. Hence, the insertion of a retrotransposon upstream of a gene provided an alternative promoter, leading to a novel transcript with new tissue-specificity [97].

A particularly well-studied and interesting example of an LTR promoter effect is the Agouti viable yellow gene in mouse, reviewed in [98]. The insertion of an intracisternal A particle (IAP), a mouse-specific ERV, upstream of the Agouti gene has provided an alternative promoter. If this promoter is unmethylated, the Agouti gene is ectopically expressed, leading to a different coat colour phenotype (Figure 1.12). If methylated, normal expression occurs and the usual phenotype is observed. This particular example also demonstrates the ability of retrotransposons to create metastable alleles: genes that create different phenotypes depending on the epigenetic state of a particular locus, in genetically identical individuals.

There are also entire gene networks that have been found to rely on retrotransposon regulatory elements. Chuong *et al.* found that the interferon pathway in humans, part of the innate immune system, relies on TFBSs derived from a particular family of ERVs. Deleting these ERVs led to a reduced immune response to viral infection. They also found evidence for a similar exaptation of ERVs

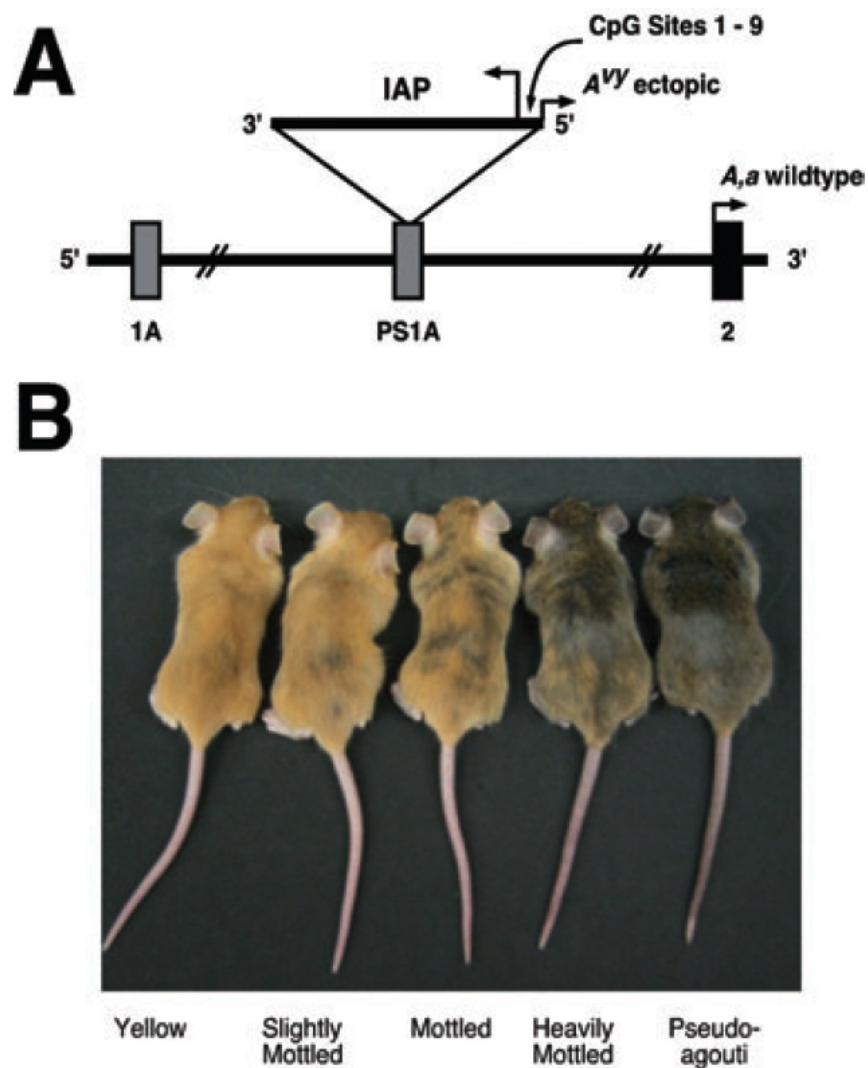


Figure 1.12: The Agouti viable yellow (*A^{vy}*) example of a retrotransposon influencing gene expression. (A) The insertion of an intracisternal A particle (IAP), a mouse-specific ERV, upstream of the *A^{vy}* gene leads to ectopic expression if the ERV is unmethylated. This leads to a metastable epiallele. (B) Genetically identical individuals with different epigenetic states leading to distinct phenotypes. Figure from [98].

in other primates [99]. A similar example is found in gene networks involved in pregnancy. In humans, progesterone triggers differentiation of endometrial cells to

form the maternal component of the placenta (the decidua). This process depends on the activation of a gene network that uses MER20 ERVs as binding sites for progesterone-responsive signalling molecules [100].

As noted above, retrotransposons can introduce new alternative splice sites, potentially leading to new isoforms of that gene. The primate-specific *Alu* SINE has been particularly well-studied in this context, with several studies showing that nearly all *Alu*-containing exons in humans are alternatively spliced [101–104]. Shen *et al.* found a particular enrichment for *Alu*-induced exons in the zinc finger proteins of primates and humans [104]. In some cases, the introduction of splice sites has led to tissue-specific isoforms, as observed in the selenoprotein N 1 (SEPN1) gene [105]. In humans, SEPN1 has a muscle-specific isoform resulting from an *Alu* exon, which is not present in macaque and chimpanzee. Lin *et al.* hypothesised that the introduction of *Alu* led to beneficial new tissue specificity [105].

Retrotransposons can also contain alternative polyadenylation sites, influencing the post-transcriptional processing of expressed isoforms [93]. The attractin (ATRN) gene in humans illustrates how a retrotransposon insertion can influence transcript diversity and function. An L1 insertion in an intron of ATRN causes cleavage and polyadenylation at an alternative site for some ATRN transcripts, while others splice out the L1. The former transcript encodes a soluble form of ATRN, whereas the latter encodes a membrane-bound protein [106].

1.2.3.2 Retrotransposons Cause Genomic Rearrangements

Retrotransposons have the potential to cause genomic rearrangements. The obvious case is the simple insertion of a retrotransposon into or near a gene. This can disrupt an open reading frame, causing the gene to lose coding ability, or separate

a gene from its enhancer or promoter. It is also possible for a retrotransposon insertion into a gene to lead to new transcripts, as described by Wheelan *et al.* They identified 15 candidate genes that had undergone an L1 insertion leading to the creation of novel transcripts, as a result of the L1 antisense promoter [107]. More recently, Criscione *et al.* used computational methods to identify 988 possible chimeric transcripts initiated from within L1 promoters. It is not known whether these transcripts are functional, and it is possible that they are transcriptional noise resulting from incomplete silencing of L1s. On the other hand, the inclusion of gene sequence could result in coding or regulatory potential.

The mechanism of retrotransposition can also cause structural variants. As discussed in detail below, retrotransposons can retrotranspose mRNA from a protein-coding gene to create a gene retrocopy. Another well-studied mechanism is transduction. L1s have a weak transcription termination signal [108,109], and so in some cases the RNA polymerase will read through this and transcribe extra sequence downstream of the retrotransposon (Figure 1.13) [108,110,111]. The whole transcript can then be picked up by the L1 retrotransposition machinery and inserted somewhere else. Hence, not only has the retrotransposon been moved, but a length of downstream sequence has also been copied to a new location. Approximately 15% of L1s in humans were found to be flanked by sequences showing evidence of transduction events [111], and L1-mediated transduction has been linked to cancer [112] and some genetic diseases [39]. Regulatory or coding elements in this extra sequence could interact with their new genetic environment, altering expression patterns or leading to new transcripts. This process could also lead to new gene copies, but there are no *in vivo* examples of L1 transduction creating new genes, other than in the context of disease. It may be that ancient examples of

this have decayed to the point that they are no longer identifiable as the result of transduction.

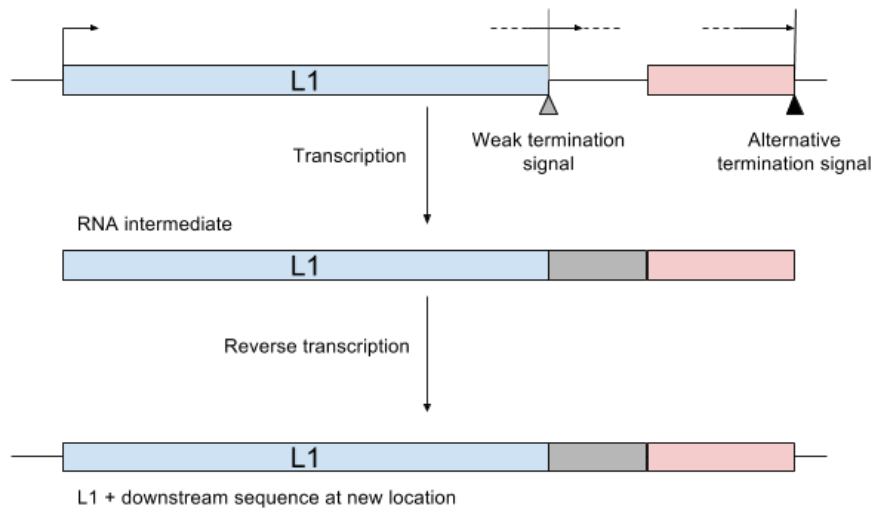


Figure 1.13: A cartoon showing L1 transduction. RNA polymerase reads through the weak termination signal in the L1 until an alternative is found. The whole transcript is then retrotranscribed and inserted into a new location. Thus, a new copy of both the L1 and neighbouring sequence is produced. This can potentially create gene copies in new locations.

The presence of multiple near-identical copies of the same sequence in the genome leads to non-allelic homologous recombination (NAHR). This can cause significant structural variants. Such events are thought to be responsible for a number of structural variants between the human and chimpanzee genomes, and have been linked to genetic diseases [39].

1.2.3.3 Retrotransposons Alter the Epigenome

Retrotransposons are thought to have played a significant role in the evolution of epigenetic mechanisms, and continue to be an important part of several epigenetic

pathways. DNA methylation, one of the most widespread and important epigenetic marks in mammals, may have evolved as a defence against retrotransposons. Similarly, piRNAs are thought to have evolved as a retrotransposon defence mechanism, but have now gained other functions in the genome.

Some investigators have observed that retrotransposons may act as nucleation points for epigenetic marks, from which marks can spread to neighbouring regions [113,114]. Conversely, L1s may be involved in X chromosome inactivation (XCI), helping genes to escape repression. During XCI, the Xist lncRNA is expressed and coats one copy of the X chromosome, silencing it. The small number of genes that escape XCI are found in low-L1 regions, and some experiments have found evidence for L1s actively nucleating heterochromatin formation [27]. However, the role of L1s in XCI, if any, is still not clear, and will require further study.

1.2.4 Transcription of Retrotransposons

Retrotransposon contributions are not limited to the genome, but extend to the transcriptome, despite the danger posed by transcriptionally active retrotransposons. There is now substantial evidence that a limited amount of retrotransposon activity is not only permitted, but essential for certain cell types. The work in this area can be separated based on the cell types investigated. The majority of work has been done in embryonic stem cells (ESCs), with some related work in induced pluripotent stem cells (iPSCs). Other studies have focused on adult somatic cells, with notable emphasis on the placenta and brain.

1.2.4.1 Retrotransposon Transcription in Pluripotent Cells

During embryonic development in humans and mice there is a general relaxation of the controls on retrotransposons, in part due to the wave of epigenetic reprogramming that occurs in this period of development. It is therefore not surprising that increased levels of retrotransposon expression are observed in embryonic cells [115, 116]. However, several studies have found evidence that specific retrotransposons are actively expressed at particular stages in development and in ESCs [117–119]. Human ERVs (HERVs) in particular seem to serve as markers for pluripotency, and may have a functional link to the pluripotent state [116, 118, 120–123]. Others have found that there is in fact dynamic control of ERVs during development, with different transcriptional profiles at different stages [124, 125].

Indeed, it appears that some retrotransposons are in fact regulated by pluripotency genes. A number of studies have shown that HERV-K and HERV-H elements are regulated by the pluripotency transcription factors (TFs) Oct4 and Nanog [122, 126]. As relatively young retrotransposons, some HERV-K elements have retained their protein-coding potential, and so their RNA can be translated to form viral-like particles in the cell [116]. It has been suggested that the presence of these particles induces an antiviral response in the embryo, protecting it from new viral infections.

1.2.4.2 Retrotransposon Transcription in Adult Cells

Several studies have shown that a significant proportion of lncRNAs contain a retrotransposon, or at least part of one [127–129]. However, the exact number


of lncRNAs present and the proportion of lncRNAs containing retrotransposon sequence vary between studies. The role of transposable elements in lncRNAs is not yet clear and the role of lncRNAs as a whole is still an area of active research. Some hypotheses have been suggested, however. Kapusta and colleagues found evidence that the presence of retrotransposons in lncRNAs allows them to form secondary structures, which may be of functional importance [128]. Johnson and Guigó have proposed that retrotransposons in lncRNAs act as sites for RNA, DNA, or protein binding [130]. Such a role may explain how functional evolution can keep pace with the rapid evolution seen in lncRNAs.

Aside from their presence within lncRNAs themselves, it is now generally agreed that retrotransposons contain a significant proportion of transcriptional start sites for lncRNAs [127, 128, 131]. Again, estimates vary as to how many lncRNAs are initiated from retrotransposons. This relationship is unsurprising, given the significant regulatory potential of retrotransposon sequences, discussed above.

A better-understood example of retrotransposon utilisation comes from placenta-specific genes. The placenta is a transient organ that mediates the exchange of gas and nutrients between fetus and mother during pregnancy [132]. In humans it also provides an environment where the “foreign” paternal antigens of the fetus are tolerated, protecting the fetus from rejection by the mother’s body [132]. Part of the placenta is formed of cytotrophoblast cells, which have the unusual ability to fuse, forming the syncytiotrophoblast [133]. This is a layer that mediates implantation of the embryo into the endometrium. It subsequently plays a number of essential roles, including mother-fetus exchange, hormone secretion, immune response regulation, and protection from pathogens. In humans, the syncytin genes syncytin-1

and syncytin-2 are have a major role in the formation of the syncytiotrophoblast. These genes are derived from human ERV (HERV) retrotransposons encoding for viral env genes [52, 134]. In their ancestral viral role, the Env proteins coded by this gene were essential for viral entry into the cell, inducing fusion of the viral envelope with the target cell membrane. This ability to induce fusion makes them useful in the development of the syncytiotrophoblast.

The syncytin-1 and syncytin-2 genes are hominoid- and primate-specific, respectively, having emerged 30 million years and 45 million years ago. Remarkably, syncytin genes derived from ERVs have been identified in diverse other mammalian species, in a fascinating example of convergent evolution. In mice, for example, the syncytin-A and syncytin-B genes are unrelated to the human genes but share an analogous function and similar characteristics, and are derived from ERVs. Homologous genes have been found in carnivores, ruminants, and other eutherian mammals, indicating that there have been independent ERV exaptation events in each of these lineages (Figure 1.14) [52].



Redacted for copyright reasons; please see Figure 4b from [52]

Figure 1.14: Syncytin genes in mammalian lineages. In a remarkable example of convergent evolution, each lineage has independently evolved syncytin genes from ERVs. Figure from [52].

Despite the various examples of functional retrotransposon transcription in

both adult and embryonic cells, the transcription of retrotransposons still has the potential to cause insertional mutations through retrotransposition. Remarkably, there is growing evidence that retrotransposition events may in fact play a useful role, specifically in cells of the neuronal lineage (reviewed in [135]). Several studies have shown that these cell types accommodate L1 activity leading to new insertions. The frequency of these insertions is disputed, with estimates varying between studies, possibly due to differences in the techniques used to measure insertional frequency. These inconsistencies aside, it is generally agreed that L1s are mobilised in the brain, and these events can be traced back to neuronal precursor cells. These insertions are enriched in genes related to neurobiology and in neuronal enhancers, and therefore have an increased chance of creating a new molecular phenotype in the brain. As yet, the function of these insertions is unclear. There is no obvious evolutionary advantage to allowing individual insertions, as they cannot be passed to subsequent generations [135]. It has been hypothesised that L1 activity in neuronal precursors gives the host organism an advantage by creating genetic mosaicism, akin to the V(D)J system in adaptive immunity. There is much work to be done before the purpose of neuronal L1 activity becomes clear, however.

1.2.5 Retrotransposon Outlook

The role of retrotransposons in the genome is not yet completely clear, but a picture is emerging showing their remarkable contribution to transcriptional and functional evolution, both directly and indirectly. Their threat to genome stability has forced the evolution of a myriad of defence mechanisms, which have

since been adopted for other functions, increasing the complexity and power of the epigenome. The regulatory sequences they contain combined with their spread throughout the genome have given host genomes the opportunity to modify existing transcriptional networks and create entirely new ones. Retrotransposition activity has shuffled the genome, creating new copies of genes, which act as raw material for the creation of new genes and the modification of existing ones. Finally, their contribution to the transcriptome, both coding and non-coding, is only just emerging. Useful protein functions have been adopted and led to the evolution of some of the most fundamental mammalian traits. As the non-coding genome begins to be understood, retrotransposons will be included in the discussion, as it is already abundantly clear that they make up a great part of it. While retrotransposons are undoubtedly dangerous and continue to lead to mutations and disease, their presence in our genome has become indispensable.

1.3 Gene Retrocopies

Throughout the mouse and human genomes, there are full and partial copies of genes. Here, I will review the mechanisms by which gene copies arise, focusing on those that arise as a result of retrotransposon activity. I will then discuss the evolutionary and functional roles they have acquired.

1.3.1 Retrocopy Origins in Mammals

New copies of existing genes can arise through either DNA-based or RNA-based mechanisms. DNA-based mechanisms include segmental duplications overlapping all or part of a gene, and large-scale genome duplication events leading to copies of

whole chromosomes [136]. RNA-based mechanisms rely on the activity of a reverse transcriptase (RT) enzyme that uses an mRNA template to synthesise DNA, which is then integrated into the genome [137]. In many eukaryotes, and particularly mammals, the major source of reverse transcriptase is retrotransposons [40, 42, 138, 139]. As such, gene copies arising in this way are often known as retrocopies. These arise when the reverse transcriptase from a retrotransposon erroneously targets mRNA from a gene, rather than retrotransposon's own RNA. The mRNA is then used to synthesise DNA, which is inserted into a new location (Figure 1.15).

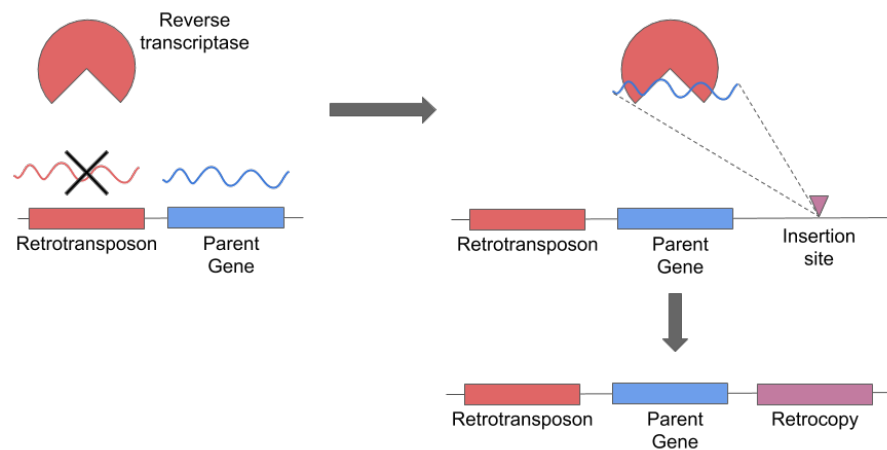


Figure 1.15: The process by which new retrocopies are formed as a result of the activity of reverse transcriptase (RT) from a retrotransposon. Instead of targeting the retrotransposon RNA, the RT enzyme targets mRNA from a gene. New DNA is then synthesised using the mRNA as a template, and this is inserted into the genome at a new location. Hence, a new partial copy of the original gene arises.

The high level of retrotransposon activity in many eukaryotic lineages has led to large numbers of retrocopies in modern genomes. In mouse, the estimated number of retrocopies ranges from 6,000 to 18,500, depending on the method used to identify them [140]. In general, retrocopies are found by searching for sequences matching mRNA from existing transcripts. This process is confounded by acquired

mutations in retrocopies and a lack of sequence motifs that consistently identify retrocopies [140].

The retrotransposons responsible for retrocopy formation vary between species, depending on which have been most active. In mammals, LINE-1 (L1) elements seem to be the most significant contributors. Studies in cultured cells have demonstrated the ability of L1 elements to create gene retrocopies using mRNA as a template [40, 42, 138, 139]. There also seems to be a correlation between L1 activity and retrogene formation. For example, the platypus seems to have relatively few retrocopies [141], commensurate with its small number of L1 elements [137], while the mouse has many thousands of retrocopies [141], and a large number of L1 elements [8]. Many recent retrocopies show the hallmarks of L1-based retrotransposition, such as target-site duplications [40, 42, 138]. These hallmarks tend to decay over time, and so it is harder to demonstrate conclusively that older retrocopies originated from L1 activity [140]. Alongside L1s, ERV retrotransposons are also abundant in mammalian (and other) genomes, and also encode for reverse transcriptase enzymes. One might expect that ERVs would then also give rise to retrocopies, but it is not yet clear whether this is the case. Introduction of ERV RT enzymes into cell lines does not give rise to new retrocopies, as L1 RT does [139]. However, Tan *et al.* recently found evidence for LTR-mediated gene duplication in a range of metazoan species, including mammals [142]. The decay of tell-tale marks that would identify the pathway responsible make it difficult to reliably assign retrocopies to the retrotransposon family responsible. It may also be that LTR-mediated duplications are sufficiently rare that experimental assays cannot detect them - indeed, fewer than 0.05% of L1 retrotransposition events led to retrocopy formation [42]. While L1s are probably the major players in retrocopy

formation in most mammals, it would be unwise to dismiss the potential impact of other autonomous retrotransposons.

The structure of the retrocopy compared to its parent gene depends primarily on the RNA molecule that was retroposed. The majority of retrocopies observed in the mouse and human genomes are partial copies, sometimes known as processed pseudogenes [137]. In these cases, the mature mRNA has been captured by the retrotransposition machinery. Therefore, the new copy will lack any untranscribed regulatory regions, introns will have been spliced out, and it will have a polyA tail. While these retrocopies appear at first to be functionally incompetent, especially as they lack regulatory sequence, there is now evidence that they make significant contributions to genome evolution and the transcriptome, discussed below. In rare cases, full retrocopies of genes have also been observed, when a non-processed mRNA has been retrotransposed. These will have retained introns from the parent gene, and will not have a polyA tail. In cases where the parent gene has multiple transcriptional start sites, a promoter may also be included in the retrocopy, potentially leading to expression of the retrocopy from its own promoter [141].

It should be noted that for a retrocopy to become fixed in a lineage, its formation must take place in the germline. Therefore, there are two basic requirements for a retrocopy to form: its parent gene must be expressed at some level in the germline, and a retrotransposon able to carry out retrotransposition must be active. This has led to a bias in the formation of retrocopies of germline-specific genes, and genes that are expressed ubiquitously across tissues [143–145].

1.3.2 Contribution of Retrocopies to Genome Evolution

Gene retrocopies make good source material for new genes and transcripts. Despite the frequent lack of promoters and introns, there are several examples of retrocopies evolving into new genes, often with a different function to their parent. In general, retrocopies tend to evolve distinct expression patterns from their parents, becoming involved in new pathways and thus evolving new functions [137,140]. In some cases, the retrocopy is still able to encode for a protein, but mutations lead to a change in subcellular localisation of the new protein [146]. An example of this is glutamate dehydrogenase 2 (GLUD2), which emerged in a common ancestor of apes and humans as a retrocopy of GLUD1 [147]. GLUD1 localises to mitochondria and the cytoplasm, but a substitution in the protein’s sequence led to GLUD2 targeting only mitochondria [148,149]. It has been suggested that this change contributed to adaptation of GLUD2 for a new function in metabolism in the brain [149].

Aside from the creation of new genes, retrocopies inserted into or near an existing gene can alter the expression patterns of the existing gene, or lead to the creation of novel transcripts. Illustrative examples of these “fusion genes” are found in owl monkeys and macaques. Remarkably, both examples involve the same pair of genes, tripartite motif containing 5 (TRIM5) and cyclophilin A (CypA). TRIM5 is an antiviral defence gene, while CypA binds strongly to retroviral capsids [137]. In both cases, a retrocopy of CypA inserted into TRIM5, replacing the capsid-binding domain of TRIM5 and leading to a more effective antiviral defence protein [150–152]. In owl monkeys, the CypA retrocopy inserted into an intron of TRIM5, and alternative splicing led to the new TRIM5-CypA fusion transcript [150,152]. In macaques, CypA inserted into the 3’ UTR of TRIM5,

again leading to a novel fusion transcript [151, 152]. While two cases of the same fusion gene arising independently may seem unlikely, CypA is highly expressed in the germline, increasing its chances of being retroposed [137], and the creation of an improved antiviral gene will have increased its chances of being selected for.

The above two examples illustrate the potential for retrocopies to give rise to new transcripts with novel functions, and a number of other specific examples have been characterised (reviewed in [137] and [140]). It should be noted, however, that these are the exceptions, rather than the rule. Even the most conservative estimates of retrocopies in primates and mouse put the numbers in the thousands, and it is certainly not the case that every single one of these has or will give rise to a novel coding gene. While they have undoubtedly played an important role in gene evolution, this may not be their most significant contribution to the transcriptome.

1.3.3 Retrocopy Expression and the Germline

A consistent observation in studies of retrocopies is that they are frequently expressed in the testes [141, 153, 154]. It has been suggested that the permissive transcriptional environment of the testes allows transcription of DNA that is usually silent, including retrocopies. In this way, retrocopies start off expressed in the testes and gain a useful function, in some cases leading to expression elsewhere and the evolution of new genes [137]. An alternative hypothesis is that retrocopies preferentially insert into regions of open chromatin. In testes, this would mean retrocopy insertions near to germline-expressed genes, and so the retrocopies are therefore more likely to be transcribed when these nearby genes

are transcribed [137]. These hypotheses are not mutually exclusive, and different retrocopies may have become expressed for either of these reasons.

It has also been consistently observed in both mammals and *Drosophila* that a disproportionate number of functional retrocopies originate from genes on the X chromosome [141, 155–157]. In mammals, these retrocopies are expressed in the testes during and after meiosis [141]. During this period, their parent genes are silenced as a result of meiotic sex chromosome inactivation (MSCI), when the X and Y chromosomes are segregated and transcriptionally silenced [158]. Based on these observations, the “out of X” hypothesis states that retrocopies originating in the X chromosome are selected for in order to compensate for silenced parent genes that are needed during meiosis [141].

1.3.4 Non-coding Transcription of Retrocopies

As noted above, relatively few retrocopies have been characterised as having evolved into protein-coding genes, compared to the number of retrocopies identified. However, several studies have found widespread transcription of retrocopies [141, 153, 154, 159–161]. Rather than being protein-coding, these transcripts contribute to the repertoire of non-coding RNAs (ncRNAs). In particular, the sequence similarity between a retrocopy and its parent transcript can lead to interesting and functionally relevant interactions between the two, at the transcriptional and post-transcriptional levels. Here I discuss four such mechanisms.

1.3.4.1 Targeting of Epigenetic Modifications

At the transcriptional level, there is evidence to suggest that retrocopy RNA can alter chromatin state at the parent locus. Phosphatase and tensin homolog (PTEN) is a tumour-suppressor gene in humans with a retrocopy, PTENpg1, that expresses three antisense lncRNA isoforms [162]. These RNAs are complementary to the parent RNA from which they originated. Johnsson *et al.* found that suppression of one of these transcripts in human cell lines resulted in increased PTEN expression, and that this was mediated by chromatin remodelling complexes [162]. They hypothesised that the retrocopy RNA targets the chromatin modifiers to the parent locus. Similar mechanisms have been observed in antisense lncRNAs (e.g., HOTAIR [163]), and it may be the case that numerous retrocopy transcripts act in a similar way.

1.3.4.2 Retrocopy LncRNAs as Micro RNA Decoys

Micro RNAs (miRNAs) are a class of short (around 23bp) non-coding RNAs that mediate post-transcriptional gene regulation (reviewed in [164]). This is achieved by binding of a miRNA to a target mRNA through sequence complementarity. This inhibits translation, either by repressing translation directly or by triggering RNA degradation [164]. Any RNA with the correct recognition site could be targeted by the relevant miRNAs, and as such retrocopy transcripts with intact miRNA recognition sites could act as decoys for their parent mRNA. Several studies have shown this occurring in genes with retrocopies, including PTEN, in the context of cancers [165–169]. In these cases, the retrocopy transcript is bound by miRNAs and so the parent mRNA is not repressed.

1.3.4.3 Retrocopy Antisense LncRNAs

In order to act as a miRNA decoy, the retrocopy transcript must have sequence matching that of the parent (at least at the miRNA recognition site). Such a mechanism might be thought to account for the observed retrocopy transcription. However, several recent studies have shown that retrocopies are also antisense transcribed, giving rise to lncRNAs complementary to the parent [159–161]. These have the potential to form RNA:RNA duplexes with their parent mRNAs, which can lead to both up- and down-regulation of the parent [170, 171], but it is still unclear to what extent this occurs and what its functional consequences are.

Korneev *et al.* provided an early demonstration of this mechanism using the central nervous system of the *Lymnaea stagnalis* snail as a model. They found that transcripts from a nitric oxide synthase (NOS) retrocopy form RNA:RNA duplexes with NOS mRNA *in vivo*. This prevents translation, leading to a reduction in NOS enzyme activity [172].

More recently, genome-wide studies have found examples of retrocopy antisense transcripts in human, but estimates differ between studies. Muro and Andrade-Navarro found 87 human retrocopies with antisense transcripts [159], whereas Bryzghalov *et al.* found only 35 [160]. The latter study also found expression correlation for three retrocopy RNA/parent RNA pairs (two positive, one negative), and ten pairs with high sequence identity. While these few serve as interesting examples to pursue, they do not provide a genome-wide picture. Milligan *et al.* performed a more comprehensive analysis in humans, comparing lncRNA databases with retrocopy annotations. They found 313 potential antisense-expressed retrocopies, but did not investigate the effects of these on the expression of their parent

genes [161].

1.3.4.4 Retrocopy Short RNAs

As well as interacting with short ncRNAs, retrocopies can be a source of short ncRNA, including miRNA, piRNA, and endogenous short interfering RNA (endo-siRNA), although it remains to be seen whether this is a major role for retrocopies.

The current evidence for retrocopy-derived miRNA is not convincing. Devor identified two primate-specific miRNA loci overlapping retrocopies [173], and other examples of miRNA/retrocopy overlaps can be found [140]. However, these account for very small proportions of miRNAs and retrocopies, and without knowledge of their regulatory targets it is difficult to assess whether this is truly an important link, or simply a handful of coincidences.

A better-studied phenomenon is the transcription of piRNAs from retrocopies. Studies in the common marmoset, human, mouse, and pig have found antisense-oriented retrocopies present within piRNA clusters [174–177]. PiRNAs transcribed from such clusters could potentially regulate parent genes via the same mechanism used to silence retrotransposons, and this has indeed been observed in mouse and pig [176, 177]. This suggests additional roles for both piRNAs, as regulators of genes as well as retrotransposons, and for retrocopies, as sources of regulatory small RNAs.

Finally, retrocopies have been found to generate endo-siRNAs, particularly in the mouse oocyte [83, 178] where they regulate protein-coding transcripts. A similar phenomenon has been observed in hepatocellular carcinoma [179].

1.3.5 Retrocopy Regulation

A question that arises from the observation that retrocopies are transcribed is how they are regulated. As mentioned above, many retrocopies do not carry with them the promoter or other regulatory elements of their parent gene, so in most cases regulatory elements must be found elsewhere if the retrocopy is to be transcribed. There are three mechanisms that seem to account for the majority of observed retrocopy transcripts:

1. Retrocopies that insert inside a new gene can become part of a fusion gene, as described above
2. An insertion near to an existing gene may be able to “piggyback” on the existing promoter, estimated to occur for around 11% of retrocopies in human and mouse [141]
3. A new promoter can evolve, either from an existing proto-promoter (e.g., part of a retrotransposon) or a CpG island; according to Carelli *et al.*, approximately 51% of retrocopies in humans and 38% in mouse have a promoter overlapping a CpG island [141]

1.3.6 Retrocopy Outlook

A number of specific and well-characterised examples have conclusively demonstrated the potential for retrocopies to evolve into novel genes and transcripts, and to alter those that already exist. Undoubtedly there are still many such cases to be discovered, and they will continue to provide insight into the evolution of genes and their functions. Still far from clear, however, is the role of retrocopies in

the non-coding transcriptome. It is clear that they do indeed contribute, forming a subset of lncRNAs and giving rise to short ncRNAs. The proportion that are transcribed, and their relationship with their parent transcripts, is not decided. Their contribution to piRNAs is becoming increasingly well-characterised, and this is made easier by the significant existing body of work on piRNAs and their mechanism. In contrast, the role of long antisense retrocopy transcripts still requires much work. Do they regulate parent transcripts through RNA:RNA duplexes? Is this a positive or negative regulation, and to what extent does it occur?

1.4 RNA Sequencing

Much of the work discussed so far was made possible by the development of high-throughput sequencing techniques. Genome sequencing revealed the significant retrotransposon and retrocopy content of the human and mouse genomes. RNA sequencing (RNA-seq) opened up the transcriptome, allowing us to produce a complete snapshot of the transcriptome in a sample, leading to the discovery of the thousands of previously unknown ncRNAs. RNA-seq datasets also form the basis of this thesis, so here I summarise the experimental procedure and bioinformatic analysis, and discuss the challenges presented by repetitive DNA.

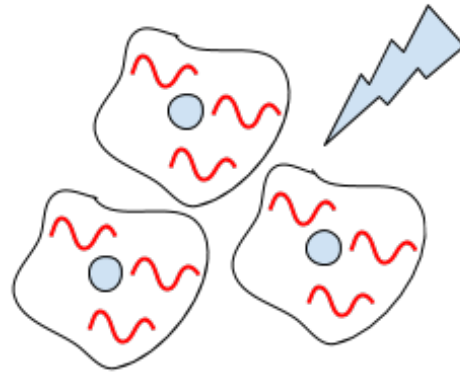
1.4.1 Summary of Experimental Procedure

There are three main steps required to produce RNA-seq reads from a biological sample: extracting and preparing the RNA, preparing a library for sequencing, and the sequencing itself. These three steps are outlined below, based on [180–182].

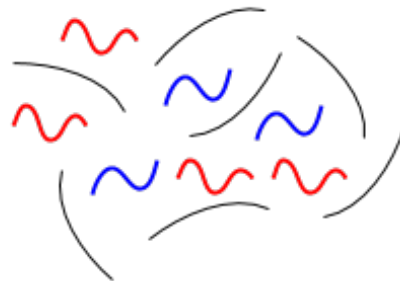
1.4.1.1 RNA Preparation

The goals of this step are to isolate and purify RNA from a biological sample, and to enrich for target classes of RNA. The procedure is summarised in Figure 1.16. Nearly all RNA-seq experiments will follow this process, and the main differences will come at the RNA filtering step (step 3 in Figure 1.16), as different experimental designs require a different subset of the RNA present in a cell. Ribosomal RNA (rRNA) accounts for more than 80% of the RNA in a cell, and so sequencing the RNA at this stage would result in rRNA accounting for nearly all of the reads and obscuring signal from other transcripts; therefore nearly all experiments will remove rRNA. This can be achieved by selectively removing rRNA, or by directly selecting for another class of RNA, such as mature mRNA. Experiments focusing on smaller RNA classes, such as miRNA, siRNA, or piRNA, can use size selection techniques to enrich for their target class, as these are much smaller than most mRNA and rRNA.

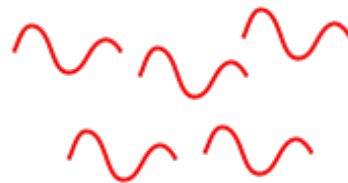
1. Disrupt cells in sample (chemical and mechanical means) to obtain cell lysate



2. Isolate RNA from cell lysate, keeping the integrity of the RNA molecules intact



3. Filter RNA molecules, e.g., remove ribosomal RNA



4. RNA ready for library preparation

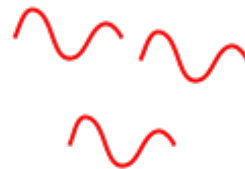


Figure 1.16: The major steps involved in extracting RNA from a sample, as the first step in RNA-seq.

1.4.1.2 Library Preparation

Once the RNA has been isolated, purified, and target-enriched, it must be converted into double-stranded complementary DNA (cDNA) for sequencing. In addition, most modern sequencing technologies require platform-specific adapters to be added at either end of the DNA molecule. These steps are summarised in Figure 1.17.

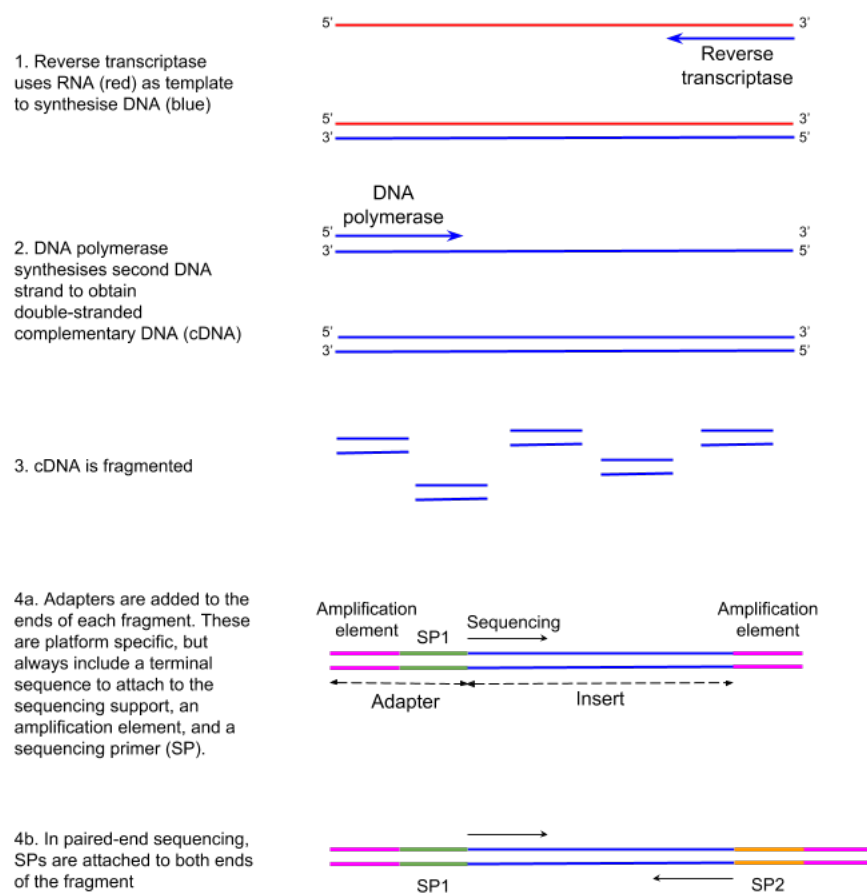


Figure 1.17: The major steps involved in preparing a library for RNA-seq, starting from purified RNA. Partially adapted from [180].

As shown in Figure 1.17, sequencing platform-specific adapters must be at-

tached to the ends of the fragments before they can be sequenced. The sequence between the adapters, known as the insert, is often longer than the maximum length that can be sequenced by the platform. By including a sequencing primer at both ends, the insert can be sequenced from both ends, resulting in a pair of reads. Knowing that a given pair of reads come from the same insert improves the efficiency with which reads can be aligned to a reference after sequencing.

Additional components can be included in the adapters, such as barcode sequences. These sequences identify which sample a given read originated from, allowing multiple samples to be sequenced simultaneously, which lowers the cost per sample. Protocols have also been developed that preserve the strand from which the original RNA was transcribed, improving mapping efficiency and enabling analysis of the direction of transcription.

1.4.1.3 Sequencing

The final stage of the RNA-seq protocol is to place the prepared sample in a sequencing machine. Different companies have developed different technologies to carry out the actual sequencing. Illumina is a popular choice, and is one of the technologies based on short reads (typically between 75bp and 150bp). Emerging sequence technologies are changing the way in which HTS is approached. Pacific Bio (PacBio) sequencing, for example, produces significantly longer reads than Illumina, on the order of 10,000bp. While PacBio reads have a higher error rate than short-read sequencing, their significantly increased length makes them very useful, particularly in resolving repetitive regions. Recent genomics studies have made use of hybrid approaches, using long PacBio reads to build rough scaffolds, and filling in the details with short reads. Nanopore sequencing technologies, such

as the MinION sequencer, also produce longer reads, and at a fraction of the cost and size of current standard sequencers. While these new technologies hold great potential, many studies still rely on Illumina and similar technologies, due to their low error rate, established protocols, and the existence of a wide range of analysis tools.

1.4.2 Bioinformatic Analysis

HTS techniques produce staggering amounts of information, and would be useless without efficient and accurate computational tools to analyse them. With current sequencing techniques, this information is presented as large numbers of short reads, usually between 75 and 150 base pairs (bp) in length. These reads are short substrings of the larger molecule (either DNA or RNA) being sequenced. The main goal of many RNA-seq analysis pipeline is to use these reads to discover which transcripts were present in the sample, and how abundant these transcripts were relative to each other. The following summary is based on [181, 183].

The first step in analysing RNA-seq reads is often to map them to some kind of reference, either a transcriptome or a genome. The quickest and easiest method to produce abundance estimates is to use an existing reference transcriptome. If one is not interested in discovering new transcripts, and there is a sufficiently good reference available for the organism in question, this method is effective. This can be used for straightforward abundance estimation of known mRNA, or more involved analysis (e.g., splicing).

If a reference is not available, or the experiment aims to discover novel transcripts, the RNA-seq reads can instead be mapped to a genome. This relies on a

sufficiently good genome being available, although more and more model organisms have high quality genomes. Once the reads have been mapped to the genome, the transcripts that were present in the sample can be reconstructed. This is usually done by looking at overlapping reads and estimating the most likely set of transcripts that would explain the reads present. The possibility of reads overlapping splice junctions must also be considered. In this way, transcripts not present in a given reference transcriptome can be found.

Finally, in the absence of a reference genome, transcripts can be reconstructed *de novo*. This is similar to the *de novo* construction of a genome. Reads with shared sequence are grouped together, and the most likely transcripts are constructed using these overlaps. This is a computationally intensive process, with less reliable results, and so should be avoided in favour of one of the above methods, if possible.

In order to calculate relative abundance estimates, the relative number of reads mapping to a given transcript is used as a proxy for transcript abundance, based on the assumption that an RNA molecule is just as likely to be sequenced as any other. In order to make abundance estimates comparable between different samples, the read counts for each transcript must be normalised. The most popular normalisations are reads per kilobase per million (RPKM), fragments per kilobase per million (FPKM), and transcripts per million (TPM). While FPKM and RPKM have been used for several years, and still are, TPM is becoming increasingly popular.

It should be noted, however, that these estimates are not truly comparable between samples, and only measure the relative abundance of a given transcript within a sample. In order to compare expression between samples, several statistical methods have been developed that can estimate the differential expression of

a transcript between samples, often using read counts. These can be used with multiple replicates to find differentially expressed transcripts or genes between multiple conditions, e.g., between treatment and control.

Many abundance estimation methods rely on an alignment of reads to a transcriptome. Recently published methods, such as kallisto [184], Sailfish [185], and Salmon [186] do not require an alignment, and instead create a “pseudoalignment”. This is achieved by splitting reads into short k -mers, sequences of k nucleotides, where k is typically around 25 (although this can be varied). The number and positions of occurrences of each k -mer in the reference is then obtained, and this can be used to estimate read counts and abundance for each transcript. The exact algorithm for doing this differs between software packages. These methods have proven to have comparable accuracy to alignment-based methods, and are usually orders of magnitude faster to run.

1.4.2.1 RNA-seq and Repeats

Repetitive sequences present a significant challenge to current sequencing technologies. Nearly all known genomes have repeats to some degree. These repeats are often much longer than the usual length of a read, which creates the problem of multimapping reads: reads that map equally well to multiple locations in the genome. This can make it difficult to assemble genomes accurately, and in the context of RNA-seq, it can have a significant effect on abundance estimation and transcript reconstruction [187]. Early approaches, such as ignoring multimapping reads [188, 189], or using heuristic methods to assign multimapping reads to single loci [190] risk introducing bias into one’s analysis, even if one is not considering repeats specifically [191].

Fortunately, bioinformatics tools have now been developed that account for repetitive sequences. Alignment tools often give users the option to retain reads mapping to multiple sites. Many recently developed abundance estimation tools are based on multimapping-aware models. RSEM, for example, uses expectation maximisation to assign fractions of reads to different sites based on the number of reads uniquely mapping to that site [192]. Multimapping reads are still an issue, and will remain so while short-read sequencing technologies dominate. However, those developing RNA-seq analysis pipelines now have a range of tools available that will handle multimapping reads in a rigorous way, and can avoid repeat masking and other naive approaches.

1.5 Motivation and Open Questions

It is now clear that retrotransposons are an essential part of mammalian genomes, and have shaped their evolution. They have made and continue to make significant contributions to the transcriptome. This occurs directly, by transcription of retrotransposons, and indirectly, by altering transcript regulation, and by formation of retrocopies which are then transcribed.

In both cases, high-quality RNA-seq datasets combined with repeat-aware bioinformatics pipelines are required to accurately quantify their transcription. The BLUEPRINT Consortium Work Package 11 (WP11 - mouse models) has produced high-quality RNA-seq data from purified cell populations in both the mouse reference strain and a wild-derived strain. These are ideal for exploring the retrotransposon-derived transcriptome.

1.5.1 Aims

The direct transcription of retrotransposons has been well-studied in stem cells and the germline, but with fewer studies examining adult somatic cells. In both cases, many of these have either focused on specific retrotransposon elements or families, or have quantified their transcription by mapping reads to retrotransposons. Several studies have examined the relationship between lncRNAs and retrotransposons, finding that retrotransposons contribute to a majority of lncRNAs, but do not explore the transcripts in detail. It is now well-established that lncRNAs display cell-type specificity, but to what extent is the retrotransposon content cell-type specific? I therefore aim to:

1. Produce a comprehensive catalogue of retrotransposon content in murine somatic cell transcriptomes using a repeat-aware bioinformatics pipeline
2. Test whether retrotransposon expression displays cell-type specificity by comparing B and T lymphocytes
3. Test whether retrotransposon transcripts affect expression of protein coding genes in *cis*

Understanding how retrotransposons contribute to transcripts is essential for understanding why they contribute to transcripts, which in turn will shed light on the possible functions of these transcripts. At present, it is known that many thousands of individual lncRNAs have been identified, but their function is unclear, either individually or as a whole. A clear and precise picture of their retrotransposon content will be important for understanding this.

It is becoming increasingly clear that retrocopies contribute to the transcriptome, either by altering existing transcripts or giving rise to new ones. In some cases, retrocopies have evolved to form new genes, but many have remained as partial copies lacking coding potential. Nonetheless, their expression has been observed in multiple species and multiple cell types. The function of these transcripts, if any, is not clear, although there is evidence that retrocopy transcripts are involved in regulation of their parent transcripts. However, a genome-wide assessment of their impact on parent genes has not yet been published, to my knowledge, and it is still unclear which pathways they might operate through, although several have been proposed. I therefore aim to:

1. Produce a reliable catalogue of retrocopy transcription in mouse B and T cells
2. Test whether retrocopy transcripts affect the expression of their parent genes
3. Test whether any such effect takes place via a mechanism based on sequence similarity

The formation of retrocopies may simply be a side-effect of retrotransposon activity, but there are multiple examples of these seemingly random events becoming essential in gene expression and regulation. Retrocopies are certainly a rich source of raw material for gene formation, but may also be the basis for essential regulation of transcription and translation.

Chapter 2

Datasets

2.1 BLUEPRINT Datasets

The BLUEPRINT Consortium is an EU-wide scientific consortium that aims to generate multiple normal and cancer epigenomes. Purified homogeneous populations of *ex vivo* cells are being used for the analysis since, within a tissue, different cell types may have different epigenomes. These epigenomes are being studied in order to understand the role of epigenetics in health and disease, and in particular in haematopoietic cells. BLUEPRINT formed the major part of the EU's contribution to the International Human Epigenome Consortium (IHEC). BLUEPRINT is split into eighteen Work Packages, each tackling a different aspect of the larger project. The work presented in this thesis was largely conducted on data generated as part of Work Package 11 (WP11), led by Anne Ferguson-Smith and David J. Adams (Wellcome Trust Sanger Institute), which uses the mouse as a model organism. It should be noted that the other BLUEPRINT Work Packages focus exclusively on data from humans.

As part of WP11, a large number of sequencing datasets were generated, covering the transcriptome and several aspects of the epigenome. For each type of experiment, multiple samples were used to cover different mouse strains, both sexes, and two cell types. Experimental work, including cell purification; RNA, DNA, and chromatin isolation; and library preparation, was conducted by Dr Marcela Sjöberg-Herrera in the Ferguson-Smith and Adams labs.

2.1.1 Mouse Strains

Two mouse strains were used in WP11: the inbred C57BL/6J (BL6) strain, which is the mouse reference strain, and the wild-derived CAST/Eij (CAST) strain. The CAST strain is genetically distinct from BL6, with approximately 17.5 million SNPS, 2.5 million indels, and 86,000 structural variants, which are all higher than many other commonly used mouse strains [193]. In addition to the pure strains, reciprocal hybrids were also generated, but these data are not used in this work.

2.1.2 Sexes

In each strain, both male and female mice were used.

2.1.3 Cell Types

Two cell types were used for each strain/sex combination: naïve B lymphocytes and naïve CD4⁺ T lymphocytes. Naïve lymphocytes, which have not yet been activated in response to an antigen, are quiescent but still transcriptionally active. Purified populations of cells were generated using the marker properties of CD4⁺, CD25⁻, CD62L^h, and CD44^l to purify naïve T cells and CD19⁺, CD43⁻, and B220⁺

naïve B cells. The quiescent state of the cells, which places them all in the same stage of the cell cycle, further reduced the potential noise in the sequencing output. This design minimised heterogeneity within the purified cell types.

There are therefore eight possible strain/sex/cell type combinations relevant to this work (Figure 2.1). For each combination, the following sequencing experiments were carried out, with multiple biological replicates in each, as noted. In each case, 100bp paired end reads were generated.

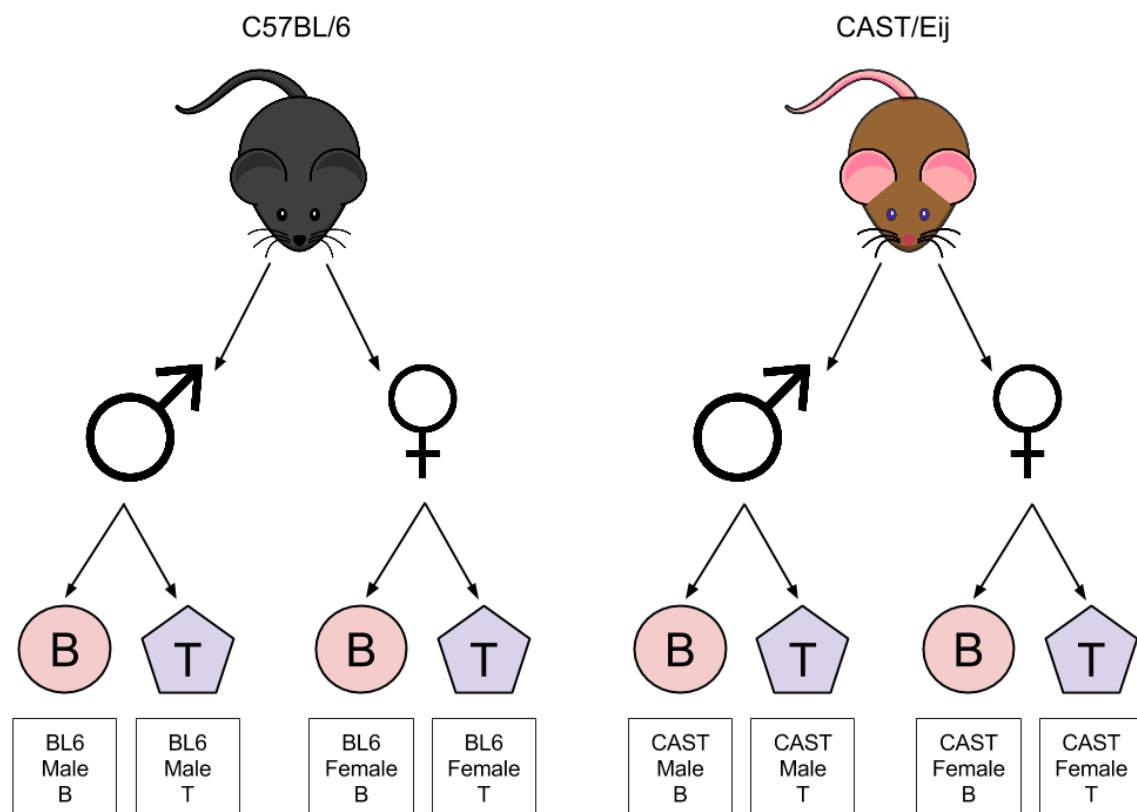


Figure 2.1: The strain/sex/cell type combinations used in the BLUEPRINT WP11 datasets.

2.1.4 RNA Sequencing

Whole-transcriptome RNA sequencing (RNA-seq) datasets used a stranded protocol and were generated by Dr Marcela Sjöberg-Herrera. Samples were depleted for ribosomal RNA. There are 12 BL6 and 11 CAST samples.

2.1.5 Whole Genome Bisulphite Sequencing

Bisulphite sequencing is used to identify which cytosines are methylated at base resolution. Whole genome bisulphite sequencing (WGBS) can be used to quantify methylation on a genome wide scale. During WGBS, the DNA in the sample is treated with bisulfite (hydrogen sulfite, HSO_3^-). This converts unmethylated cytosine to uracil, but methylcytosine is left unaffected. The sample is then sequenced. Providing one can differentiate between single nucleotide variants and genuine conversions, one can determine whether a given read covering a CpG site came from DNA that was methylated or not. By looking at multiple reads covering a CpG site, one can quantify the level of methylation at that site as the proportion of reads that are methylated. This should reflect the proportion of cells in the sample where that CpG site is methylated.

For each strain/sex/cell type combination, eight biological replicates were sequenced. The raw samples were analysed by Dr Nic Walker to produce methylation estimates at each sequenced CpG site. NW also merged replicates to improve sequencing depth and coverage, and I used merged samples for my analysis.

In addition, I applied smoothing to the methylation values, following the protocol described by Hansen *et al.* using the BSmooth software [194]. Methylation is spatially correlated between CpG sites, so CpGs near to each other tend to

have similar methylation values. Using this, the authors of BSmooth created an algorithm that adjusts methylation levels according to the levels of nearby CpG sites, weighted by their sequencing depth. In this way, the accuracy of methylation estimates is improved.

2.1.6 Chromatin Immunoprecipitation Sequencing

Chromatin immunoprecipitation sequencing (ChIP-seq) is a technique used to identify regions where a given protein is bound to DNA, such as a transcription factor. ChIP-seq comprises two main steps: the chromatin immunoprecipitation (ChIP), followed by high-throughput sequencing. During ChIP, DNA and any proteins bound to it are cross-linked. The DNA is then sheared into fragments, and bead-attached antibodies are added that are specific to the protein or protein modification of interest. The beads can then be precipitated out, the proteins unlinked, and the DNA purified. In this way, one selects or enriches for regions of DNA that had the protein of interest bound to them. The purified DNA is then sequenced in the usual way. The reads obtained from sequencing can then be analysed with the aim of finding “peaks”: regions where a significant proportion of reads map, indicating enrichment of the protein or modification of interest at that region.

In the WP11 datasets, ChIP-seq was used to identify regions with six histone modifications: H3K27ac, H3K27me3, H3K36me3, H3K4me3, H3K9me, and H4K20me3. The raw reads were analysed by Dr Hui Shi to find peaks; during this process, the latter two marks were deemed to have been unsuccessful, despite repeated attempts, and so only the first four are used in this project. Unless

stated otherwise, merged sets of peaks were used in my analysis, obtained using the bedtools merge software [195].

2.2 ENCODE Data

To extend the investigations of cell-type specificity in retrotransposon and retro-copy transcription, I downloaded publicly available RNA-seq datasets from the ENCODE project. I used data from experiment ENCSR000AJU [196], which has the following properties:

- Stranded total RNA-seq from adult BL6 liver
- 101bp paired-end reads
- rRNA depleted
- Two biological replicates, no technical replicates

These samples were generated by the Gingeras lab at Cold Spring Harbor Laboratories. While the number of replicates would ideally be larger, this was one of the few available datasets with the same technical properties as the BLUEPRINT datasets, making it a good comparison.

2.3 Proteomes

Proteomes were generated by Dr Sudhakaran Prabakaran and analysed by Dr Sudhakaran Prabakaran, Dr Ruchi Chauhan, and Ms. Chaitanya Erady (Prabakaran Group, Department of Genetics, University of Cambridge). Data were generated

for BL6 males and females, in both B and T cells, to match the BLUEPRINT RNA-seq datasets. I was provided with abundance values in each sex/cell combination for 4,030 proteins, along with additional information. An additional dataset of 1,937 proteins was also provided, in which proteins had been filtered based on false discovery rate.

2.4 Reference Genomes and Annotations

2.4.1 Reference Genomes

The majority of the analysis in this project was done on data from the C57BL/6J mouse strain, which is the mouse reference genome strain. I used genome version mm10 (GRCm38). For the CAST/Eij analysis, I used the genome assembly created by the Wellcome Trust Sanger Institute Mouse Genome Project [197].

2.4.2 Reference Transcriptome

For the BL6 analysis, I used the Ensembl GRCm38.84 annotation, released in March 2016 [198,199]. New versions have subsequently been released, but for the purposes of consistency I have used this release throughout.

2.4.3 Repeat Annotation

For all of the retrotransposon analysis in this project, I used the annotation of repetitive regions created by RepeatMasker [8] for the UCSC Genome Browser [200]. RepeatMasker is a set of software tools designed to screen DNA for interspersed repeats and low complexity regions. The output is an annotation of the repetitive

regions in the query sequence, and, if desired, a copy of the query sequence with the repeats “masked”: either converted to Ns or to lower-case characters. For some commonly-used genomes, such as human and mouse, ready-made repeat annotations are available. These annotations include the location and type of repeat, and summary statistics about its identification. In addition, recent updates from RepeatMasker also include information about how fragments of repeats, particularly retrotransposons, may be related. For example, if an ERV has had a SINE inserted in the middle of it, the now separated pieces of the ERV would be represented as a single element in two pieces.

RepeatMasker was originally developed by Smit *et al.* in the 1990s [201]. The underlying method for identifying repeats has not changed radically since then. RepeatMasker takes a set of reference repeat sequences, and searches for matching sequences in the query. For the ready-made annotations available for mouse and human, the reference sequences are based on two libraries of consensus repeat sequences:

- Repbase [202]: a library of manually curated submissions from researches, maintained by the Genetic Information Research Institute (GIRI)
- Dfam [203]: a more recent library that uses hidden Markov models to identify consensus sequences

Given a consensus library, RepeatMasker can use one of several tools to perform the search step, depending on the relative importance of factors such as speed and precision.

There are several advantages to using RepeatMasker. It is actively maintained and updated, and represents many years of expertise in the field. It is also popular

amongst groups working in the field, so comparisons with other studies are more straightforward, especially given the established nomenclature and categorisation used by RepeatMasker. The integration with the UCSC Genome Browser means that visualisation and comparison with other annotations (e.g., reference transcriptomes) is easy. In general, the availability of a ready-made and high quality annotation saves the significant time and resources required to produce one. Repbase, the repeat reference library used by RepeatMasker, is manually curated and represents the most comprehensive library of repeats available, and many years of experience in the field.

However, there are also problems associated with RepeatMasker. While it is maintained, it is not always up to date with the latest reference sequences. Similarly, while the UCSC Genome Browser advertises itself as having the most recent RepeatMasker releases, this is not always the case. For example, the most recent Repbase update was in 2015, while the most recent mm10 repeat annotation available on RepeatMasker was created in 2013. Similarly, the currently available RepeatMasker annotation on UCSC Genome Browser was created in 2012, but the most recent data on the RepeatMasker website is labelled as 2013-04-22.

The use of Repbase may affect RepeatMasker results. While manual curation can be advantageous, as it leverages human intelligence and expertise, it can also lead to biases, as acknowledged by the maintainers of Repbase [204]. (Although the Repbase maintainers have taken steps to reduce unintended bias in the submissions to Repbase.) In addition, Repbase data and methods are not openly available, and so it is difficult to assess their methods in comparison with others.

The sensitivity of RepeatMasker is difficult to assess, as there is not yet a standard set of benchmarks for repeat annotation [205,206]. There are now many tools

designed to identify and annotate repeats, many of which are specific to particular species or types of repeat [206], further increasing the complexity of meaningful comparisons. While these may be able to identify repeats with high specificity and sensitivity, they cannot be categorised without the use of a reference library such as Repbase. Aside from Repbase, few reference libraries exist. Dfam, for example, can be used, and it has been incorporated into RepeatMasker. However, the methodology used for Dfam suggests that in fact they use Repbase and RepeatMasker to produce their library, and so their results may be influenced by the same biases. In addition, *de novo* repeat annotation is a computationally intensive and time-consuming process.

I decided to use the RepeatMasker annotation available on the UCSC Genome Browser, and the work in this thesis was carried out using the version available there as of December 2015. This decision was motivated by the ease of use of RepeatMasker and the expertise behind it. In addition, being able to quickly visualise the repetitive regions alongside my own datasets proved extremely useful. The recent inclusion of fragment-joining information was also extremely useful in accurately quantifying retrotransposon transcription. At the time of writing, more recent versions of the RepeatMasker mouse annotation have become available, and it would be of interest to repeat the analysis described here with the new annotation. It would also be advisable to experiment with other repeat identification software and compare the results. More accurate retrotransposon identification should reduce noise and clarify the existing results.

2.4.4 Retrocopy Annotation

For the retrocopy analysis, I used the “Retroposed Genes V6” dataset, which is the retrocopy annotation currently displayed by default on the UCSC Genome Browser. This annotation was created using the retroFinder program [207], which finds alignments between known mRNAs and the corresponding reference genome. The resulting alignments are then filtered based on features indicative of retrocopy formation, such as poly(A) tail length and proportion of mRNA present in the alignment. The original retroFinder publication assessed this method’s efficacy and found it to be in good agreement with other existing studies at the time. Since then, the resulting retrocopy annotation has been updated to include more recent reference mRNAs.

My decision to use this annotation was influenced by several factors. Using ready-made annotations such as this leverages domain expertise, and avoids the need to recreate the annotation, which could be time consuming and may not produce comparable results. Its integration with the UCSC Genome Browser allows rapid visualisation of retrocopies in combination with other datasets, such as reference genes and retrotransposons. In particular, this annotation lists the parent reference gene corresponding to the retrocopy, where possible, using Ensembl identifiers, which enables integration with a high-quality gene annotation for analysis of retrocopy parents. It also includes detailed information on the alignment between retrocopy and parent.

As with other annotations available through the UCSC Genome Browser, it receives updates to reflect new reference data. However, these do not always reflect the most recent data available. The annotation currently available on the UCSC

Genome Browser was created in 2015 using the RefSeq gene annotation [208]; however, since then there have been multiple updates to the RefSeq library. Ideally, these changes should be used to update the retrocopy annotation, although the changes may be relatively minor, and without significant effects on the results of retrocopy discovery.

As noted in [140], the UCSC Genome Browser annotation is less conservative than other available annotations in mouse. Without a ground truth dataset it is difficult to compare these annotations. The differences between annotations likely reflect the methods and reference genes used, and choices by individual authors on what should or should not be included. By using a less conservative annotation, I aim to be as inclusive as possible. Applying the same analysis to other annotations would hopefully produce similar results, and such a comparison between the results would be of interest in future work. Using a merged dataset based on multiple annotations could be a reliable alternative.

Chapter 3

Methods

3.1 Transcriptome Reconstruction Pipeline

The aims of this project were to explore the retrotransposon and retrocopy content of somatic cell transcriptomes. The bioinformatics analysis pipeline used therefore had to satisfy the following key requirements:

- Multimapping reads preserved, not discarded or collapsed to one locus
- Novel transcripts discovered
- Multimapping reads treated properly during transcript abundance estimation

Alongside these specific requirements, the pipeline needed the usual desired characteristics, including high quality results, ease of use, and reasonable running time. The final outcome of this pipeline would be a set of reconstructed transcripts with accurate abundance estimates.

3.1.1 Comparison of Available Tools

Implementing this pipeline would require four main bioinformatics tools: a quality control tool; an aligner; a transcript constructor; and an abundance estimator. In addition, ribosomal RNA reads would have to be removed, either before or after alignment. Multiple tools have been published to accomplish each of these tasks, and so it was necessary to compare those available and choose the most appropriate for this project.

3.1.1.1 Quality Control

It is common practice to check raw reads for various quality metrics, such as phred score, adapter contamination, and GC content. (Phred score measures the likelihood that a particular base in a read is correct; higher scores mean the base is more likely to be correct.) In the event of poor quality or contaminated reads being discovered, adapters and bases with low phred scores can be trimmed to produce shorter reads with higher quality scores. This can improve the results of subsequent analysis steps.

The choice of quality control tool does not impact the results of RNA-seq analysis results directly, but it is important that the tool used provides a comprehensive set of checks with easy to interpret results. I therefore chose to use FastQC [209] to check read quality. FastQC is a popular tool that performs a wide set of quality checks and outputs the results in an easy-to-read HTML file.

Based on the results of FastQC, it appeared that some samples might benefit from read trimming. Many tools are available, each employing a different algorithm, which impacts both results and run time. There was no evidence of adapter

contamination from FastQC, so I focused on choosing a tool to trim reads based on phred score. I ran preliminary tests on cutadapt [210], ERNE-FILTER [211], and trimmomatic [212] to check their impact on FastQC output and alignment score. I discovered that trimming reads appeared to have little impact on either of these outcomes. In addition, literature on this subject suggested that trimming should be used with caution in RNA-seq analyses, lest useful information be lost [213]. Many recent aligners automatically take quality score into account during the alignment, trimming reads if and when necessary to achieve a better alignment. I therefore decided not to carry out an explicit read trimming step, and instead to choose an aligner that would handle this internally.

3.1.1.2 Alignment and rRNA Removal

There are many alignment tools available for RNA-seq reads, and so I relied on published comparison and benchmarking studies to narrow the field to a smaller number of choices. I then made a final decision based on suitability for this project, ease of use, and speed. The comparison published by Engström *et al.* [214] showed that STAR [215], GSNAP [216], and RUM [217] all produce high-quality results in comparison to other popular aligners. The paper that presented STAR demonstrated similar results [215]. A more recent comparison by Sahraeian *et al.* has demonstrated that HISAT2 [218] may perform better than STAR [219], although the HISAT2 documentation states that it is not suitable when reads mapping to many loci need to be retained, as is the case in this project.

I carried out informal testing with these STAR, GSNAP, and RUM, and investigated their capabilities. STAR stood out from the others in terms of ease of use and documentation quality. In addition, it automatically trims reads based

on alignment quality. Most importantly, a maximum number of alignments per read can be set, and each alignment is preserved in the results. Finally, STAR runs significantly faster than any other published aligner. I therefore decided to use STAR for the alignment step.

I decided to use the RSeQC software package [220] to remove reads mapping to known rRNA regions, based on its fast runtime, ease of use, and good documentation.

3.1.1.3 Transcript Reconstruction

There are significantly fewer transcript reconstruction tools available than there are aligners. The most popular available are Scripture [221], Cufflinks [222], and, more recently, StringTie [223] (the successor to Cufflinks). Scripture is poorly documented and maintained, and is difficult to use, unlike Cufflinks and StringTie. StringTie was presented as an improved version of Cufflinks, and informal testing with both demonstrated that StringTie runs significantly faster and is easier to use. In addition, the more recent versions of StringTie handle reads from stranded RNA-seq protocols. A comparison between StringTie and other tools in the context of a complete pipeline demonstrates that StringTie does indeed produce better results [219]. I therefore decided to use StringTie as the transcript reconstruction tool.

3.1.1.4 Abundance Estimation

The key requirement for an abundance estimation tool in this project is that it handles multimapping reads correctly, as this is the stage where they can have the greatest impact. There are several tools that explicitly deal with multimap-

ping reads. Popular choices in the bioinformatics community include RSEM [192] and kallisto [184], both of which perform well according to benchmarking studies [184, 192, 219, 224]. RSEM uses an expectation-maximisation algorithm to assign fractions of reads to different loci based on the number of uniquely mapped reads at each locus. Alternatively, kallisto is one of a recent group of abundance estimators that rely on a “pseudoalignment” rather than an explicit alignment. This method results in remarkable speed-ups, without a drop in quality of results. I decided to use kallisto, based on its ease of use, remarkable speed, and benchmarking results.

3.1.2 Final Pipeline

Based on the above considerations, I decided on a final pipeline using FastQC, STAR, RSeQC, StringTie, and kallisto:

1. Align raw RNA-seq reads to the appropriate genome using STAR version 2.5.0a with the following options:
 - `--outReadsUnmapped Fastx`
 - `--outSAMattributes All`
 - `--outFilterMultimapNmax 50`
2. Remove reads mapping to rRNA regions using `split_bam.py` from RSeQC version 2.6.4, with the BL6 ribosomal annotation available on the RSeQC website [225], or the rRNA annotation I developed for CAST (see below)
3. Sort reads not mapping to rRNA using samtools version 1.3.1 [226] to produce sorted BAM files

4. Reconstruct transcriptomes for each sample from the sorted BAM files using StringTie version 1.3.3 with the following options:
 - `-f 0.05`
 - `-M 0.99`
 - `--fr`
5. Merge reconstructed transcriptomes were using the `--merge` option in StringTie to produce a consensus set of transcripts for each phenotype combination. At least 3 samples contributed to each merged transcriptome in the BLUEPRINT datasets.
6. Estimate abundances of merged transcripts in each sample using kallisto
7. For BL6 samples, compare merged transcriptomes to the Ensembl annotation using the `gffcompare` tool, which comes with StringTie

3.2 Differential Expression Analysis

I performed two analyses to find differential expression between cell types. Both were done using the StringTie and Ballgown [227] software packages, following the post-alignment steps in the protocol described in [228]. Ballgown is designed to work with StringTie output, and performs well in benchmarking studies [219,229].

3.2.1 Differential Expression of Reconstructed Transcripts

After obtaining a merged set of reconstructed transcripts (see above), I used `stringtie -eB`, which reassigns reads to each transcript and therefore obtains

counts and expression values for each transcript. This also produces tables that can be used by the Ballgown software for differential expression analysis. I applied the `ballgown::stattest` function to the samples to find differentially expressed genes between cell types, accounting for sex as a potential confounding factor. I then filtered the results, retaining those where the false discovery rate-adjusted significance value $q < 0.05$. This left a set of 11,380 transcripts with associated fold-change values, p values, and q values.

3.2.2 Differential Expression of Ensembl Transcripts

This analysis was done as for the reconstructed transcriptomes, but using the Ensembl reference annotation instead of the reconstructed transcriptome. While the reference may not be as complete a transcriptome as the reconstructed one, it is easier to use the Ensembl reference to investigate retrocopy parent expression. Retrocopies that have a known parent have an Ensembl transcript to identify it. To use the reconstructed transcriptome, one would have to reliably assign reconstructed transcripts to Ensembl references. With total RNA samples, this is not straightforward, as there may be unprocessed mRNA, splice intermediates, and other RNA products that cannot easily be accounted for. Using the Ensembl reference, 10,714 transcripts remained after filtering on q value.

3.3 Comparing Reconstructed and Ensembl Transcripts

To ascertain how the reconstructed StringTie transcriptomes compared to the Ensembl reference transcripts, I used gffcompare, which is included with StringTie. The gffcompare algorithm compares exon overlaps between the query (reconstructed) transcripts and the reference and places each reconstructed transcript into one of the classes shown in Table 3.1. For classes =, c, j, o, e, i, and x, the query transcript is assigned a corresponding reference transcript. It is possible for a query transcript to match multiple references, and vice versa.

Name	Symbol	Description
Complete match	=	Query exons match the exons of a reference perfectly
Contained	c	Query exons entirely contained in reference exons
Possible novel isoform	j	Query and reference share exons and at least one exon junction
Exon overlap	o	Query exons overlap reference exons on the same strand
Possible pre-mRNA	e	Single-exon query that partially overlaps an intron
Intronic	i	Query is fully contained in a reference intron
Antisense	x	Query exons overlap reference exons on the opposite strand
Novel	u	Unknown intergenic; query does not correspond to any reference
Possible polymerase run-on	p	
Probable false positive	s	

Table 3.1: The classes used by gffcompare for query transcripts in comparison to reference transcripts.

3.4 Ribosomal RNA in CAST

In order to apply the same alignment and transcriptome reconstruction pipeline to the CAST samples as to the BL6 samples, I required an rRNA annotation in CAST so that reads mapping to these regions could be removed. As I was unable to find a published annotation, I found the regions myself using the annotation from the BL6 analysis. This approach assumes a high level of conservation between BL6 and CAST in rRNA.

I first obtained sequence for the rRNA regions used in the BL6 analysis. I then used `blastn`, part of the BLAST+ suite [230, 231], to search for matching sequences in the CAST genome. The repetitive nature of rRNA led to multiple query regions matching the same region in CAST, and so I collapsed overlapping hits into single regions. This method identified 1,407 rRNA regions in CAST, compared to 1,564 in BL6, suggesting there may have been some regions missed. However, similar ratios of the different subunits were found (Table 3.2).

Strain	Total rRNA regions	5S		LSU-rRNA_Hsa		SSU-rRNA_Hsa	
		Number	%	Number	%	Number	%
BL6	1564	1038	73.01	481	30.75	45	2.88
CAST	1407	937	73.26	424	30.14	46	3.27

Table 3.2: The results of the CAST rRNA discovery pipeline compared to the BL6 annotation.

3.5 Epigenetic State Visualisation

As described in Datasets, I had access to both methylation and histone modification data for BL6 B and T cells. I wrote custom scripts to visualise these marks at a given set of transcripts (Figure 3.1), with an option to expand the field of view (i.e, to add a given number of base pairs at either end of the region visualised). I used these to ascertain whether there were consistently different epigenetic states between transcripts with and without expressed retrocopies.

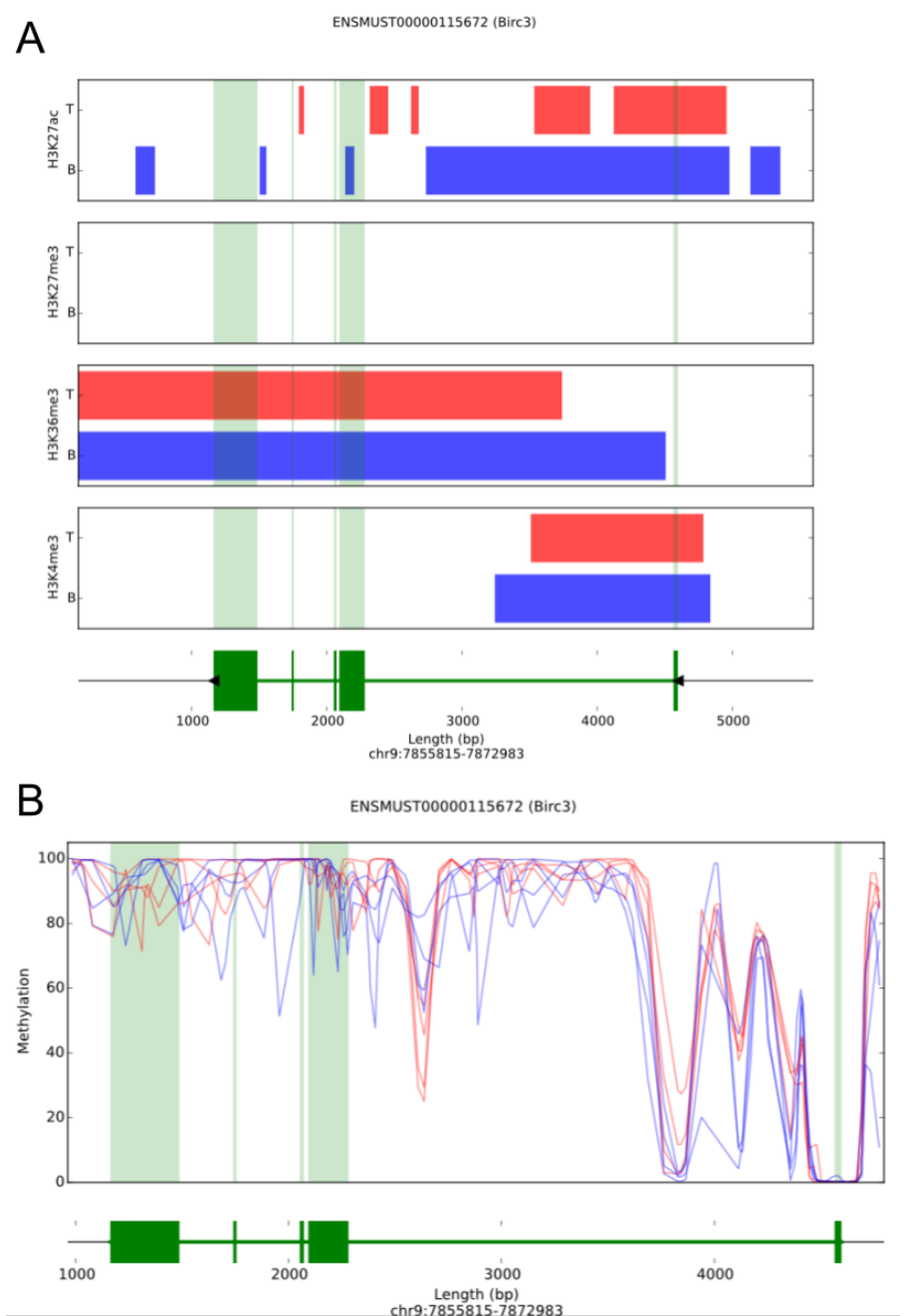


Figure 3.1: (A) An example of histone modification visualisation. Each blue/red bar represents a ChIP-seq peak. Black arrowheads show the direction of transcription. (B) An example of methylation visualisation. In both (A) and (B), blue represents B cells and red T cells. Green blocks are exons.

3.6 Proteome Normalisation

Before using the protein abundance values provided by the Prabakaran group, I normalised and transformed them in order to make the samples more comparable. I applied a median normalisation, as follows

$$v_i^j \rightarrow 100 \times \frac{v_i^j}{m^j/M} \quad (3.1)$$

where v_i^j is the i^{th} value from sample j , m_j is the median for sample j , and M is the mean of the medians from all samples. I ignored missing values. This normalisation causes all samples to have the same median. I then applied a \log_2 transformation to the non-zero normalised values. Missing values were then interpreted as zero, i.e., low abundance protein. Figure 3.2 shows the effect of normalisation on the set of 4,030 proteins.

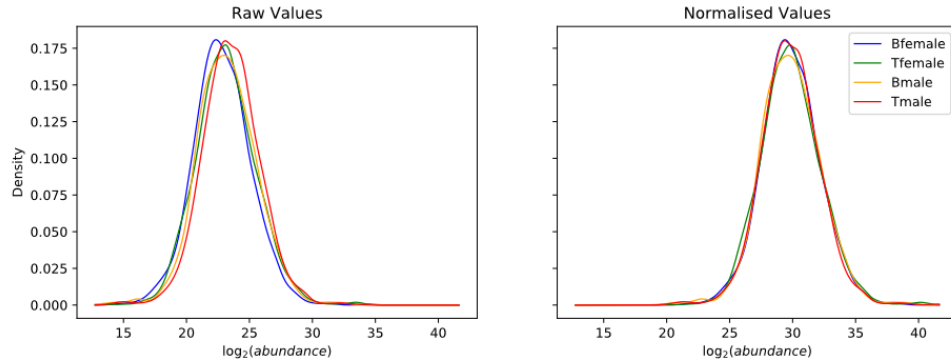


Figure 3.2: The effect of median normalisation on the distribution of protein abundance values. Missing values were ignored. The normalisation process makes the different samples more comparable.

3.7 Transcript Retrotransposon Content

Early exploration of the reconstructed transcripts and their overlap with retrotransposons showed that a naive approach to finding overlaps would result in a large number of false-positive hits. Here, the naive approach is one that uses “end-to-end” coordinates for a genomic feature (Figure 3.3) to find overlaps between two sets of features. This approach produces misleading results because it ignores the internal structure of both the transcripts and the retrotransposons. Using the beginning and end coordinates of a multi-exonic transcript will include intronic retrotransposons in the overlap results. While this may be of interest, it does not truly reflect the amount of retrotransposon sequence in the transcript, as the introns will be spliced out of the final transcript.

The retrotransposon annotation also contains internal structure. A single retrotransposon element may actually be composed of multiple blocks that are linked based on shared sequence and/or membership of the same family (Figure 3.4). These blocks may be punctuated by other types of retrotransposon, retrocopies, or other genomic features. Again, simply using the start and end coordinates to find overlaps with transcripts will give misleading results. For example, if a transcript is contained entirely between two joined L1 blocks split by an LTR, but overlaps neither of them, a naive approach will report this as a transcript high in L1 content and LTR content, when in fact it is only high in LTR content.

In order to ensure that the retrotransposon content of transcripts is accurately quantified, I developed the following method (Figure 3.3):

1. Apply `bedtools intersect` [195] to a set of exons and a set of individual retrotransposon blocks

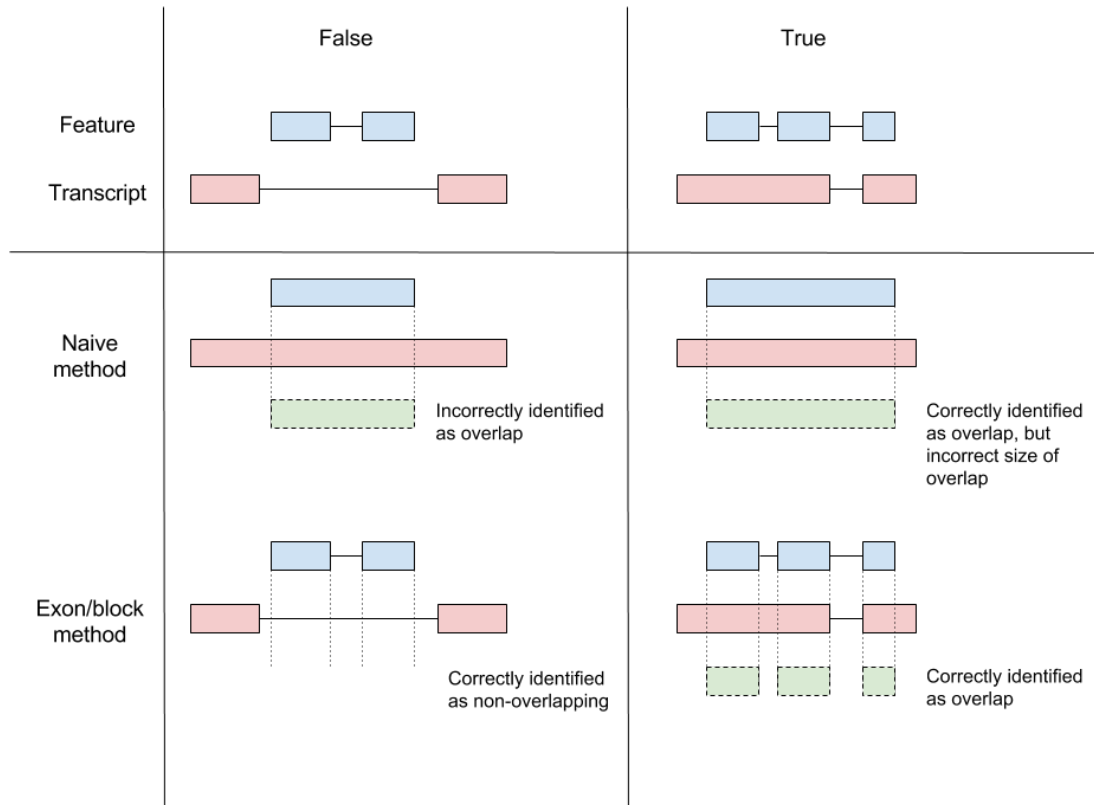


Figure 3.3: A comparison showing the potential false positives created by using a naive intersection method instead of an exon/block-aware method.

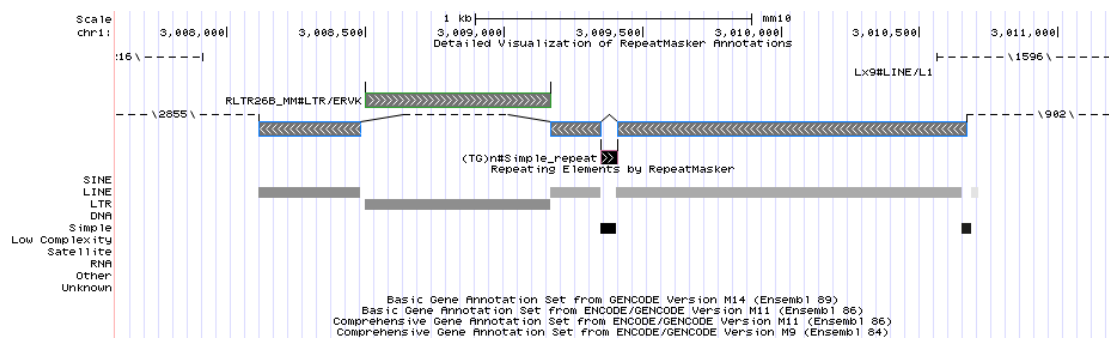


Figure 3.4: A screenshot from the UCSC Genome Browser showing the RepeatMasker tracks. A LINE element (blue) has been split by an ERV insertion (green) and a simple repeat (red).

2. Reassemble exons into transcripts and reassemble retrotransposon blocks into full elements, preserving the overlap relationships

3.8 Retrotransposon Content Visualisation

In order to summarise the retrotransposon content of transcripts in a readily understood way, I developed a barchart-like visualisation. Figure 3.5 shows an example for illustrative purposes (for analysis of these plots, see Chapter 5). The input is a set of transcripts with overlapping retrotransposons, as produced by the method described above. For each transcript, the proportion of sequence content for each retrotransposon subfamily is calculated. Non-retrotransposon sequence is labelled as unique (“UNIQ” in figures and code). Each transcript can therefore be represented as a vector of values that sum to 1, representing its retrotransposon content. The bars are coloured to represent the values in this vector; for example, a transcript with 25% LINE, 25% LTR, and 50% unique sequence would have a bar that is one quarter green, one quarter red, and half white. Agglomerative clustering is applied to these vectors in order to group the transcripts based on their retrotransposon content. The results are then plotted, with transcripts ordered based on the clusters. The bar on the left indicates the clusters, with each colour representing one cluster. (Colours are chosen on a per-plot basis based on the number of clusters, and do not indicate any kind of relationship between clusters.)

The number of clusters is chosen by maximising the silhouette score, which measures the similarity of an object to its own cluster, and therefore serves as a measure of how well matched the objects within each cluster are [232]. High values indicate that a given clustering is well-suited to the data. When using

agglomerative clustering, which structures the data hierarchically, the number of distinct clusters found is based on the distance from the root of the hierarchy. The data can therefore be divided into any number of clusters between 1 and the number of elements in the data, depending on the chosen distance from the root. In this work, I calculated the silhouette score for each value in a range of possible numbers of clusters and used the number of clusters that had the highest score.

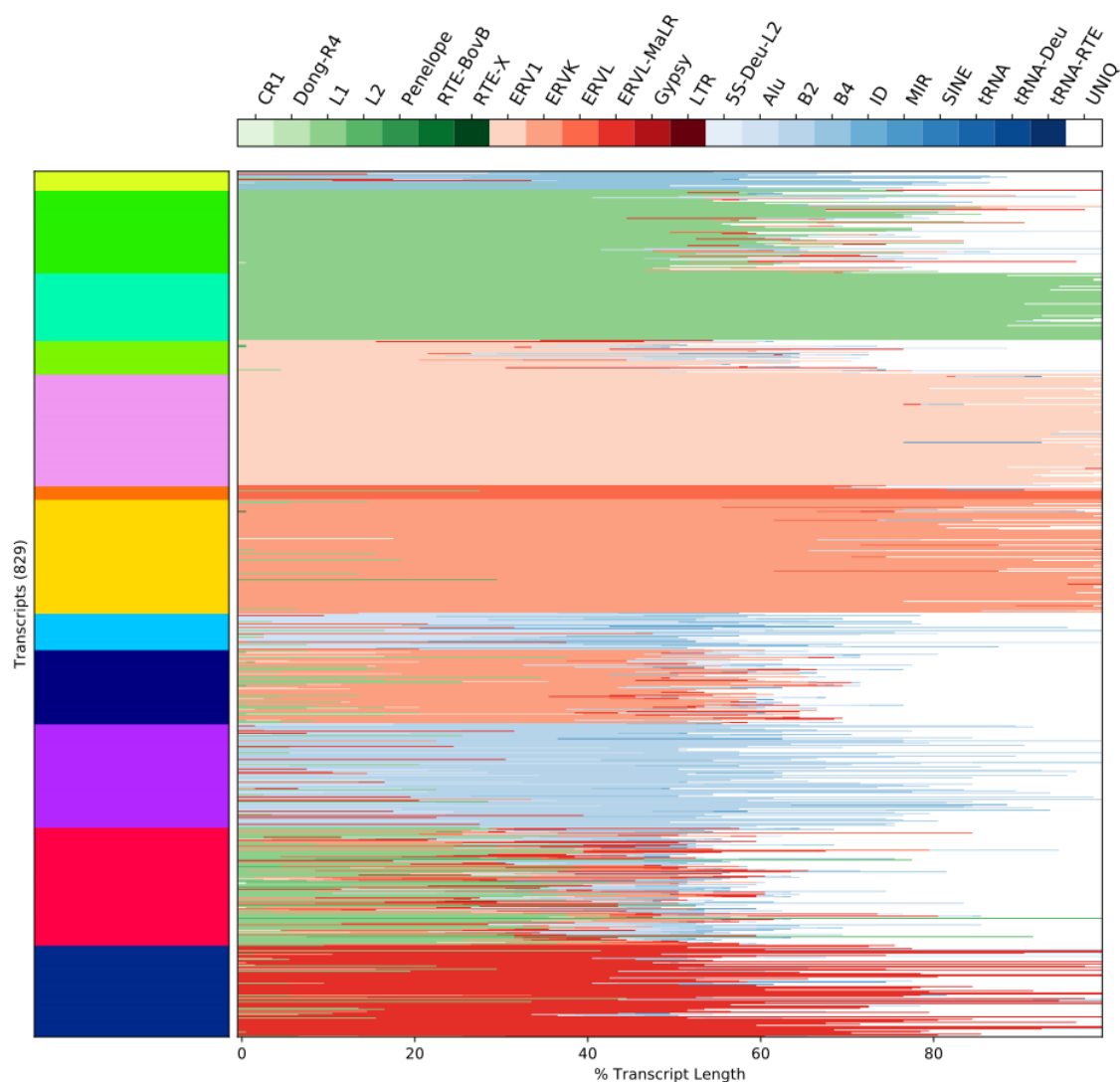


Figure 3.5: An example visualisation of the retrotransposon content of a transcriptome. Agglomerative clusters are shown in the left-hand bar. Each row in the central plot represents a single transcript. Each of the main retrotransposon families has an associated colour (LINEs, green; ERVs, red; SINEs, blue), with shades representing subfamilies, shown in the colour bar at the top. White is non-retrotransposon sequence, labelled “UNIQ”.

3.9 Cluster Comparison

Having discovered transcript clusters based on retrotransposon content, I was interested in establishing whether similar clusters existed in different samples (e.g., in different cell types). Visual inspection of the plots described above suggests that this is the case, and so I devised a method to carry out comparison of clusters in order to quantify and formalise these observations.

Separating data into clusters implies that the points within a single cluster are similar to each other; the concept of similarity is defined by the clustering algorithm used. Suppose we have two datasets D_1 and D_2 representing observations of the same phenomena under different conditions (so that $D_1, D_2 \subset D$, where D is the set of all possible observations). In the context of this project, D_1 and D_2 could be the retrotransposon content of transcripts in B and T cells.

Now suppose we cluster each dataset, so that we obtain disjoint subsets $C_i^j \subset D_i, j = 1, \dots, N_i$ where N_i is the number of clusters in D_i . Choose a pair of clusters C_1^j, C_2^k . If the elements in each of these clusters are similar in the sense defined by the clustering algorithm (i.e., they would cluster together), then these two clusters can be said to be similar. By comparing every pair of clusters C_1^j, C_2^k in this way, we can discover pairs of clusters that are similar.

To formalise this, we create a new dataset $E = D_1 \cup D_2$, and apply the clustering algorithm to create clusters $C_E^j, j = 1, \dots, N_E$. For each cluster C_1^i , we can calculate what proportion of its elements are assigned to each C_E^j . In this way we can form a matrix Q where each element Q_{ij} is the proportion of elements from C_1^i that are assigned to C_E^j . Similarly, we can form a matrix R where R_{ij} is the

proportion of elements from C_2^i that are assigned to C_E^j . Now define the matrix

$$M = QR^T$$

Therefore the element M_{ij} represents the probabilities that an element from cluster C_1^i and an element from cluster C_2^j are found in the same cluster C_E^k , summed over all clusters in E . M_{ij} can be used as a score to measure how similar two clusters are. If two clusters have elements that are often found in the same cluster C_E^k , then they will have a high score; if their elements are rarely found in the same cluster, the score will be low.

However, these scores can have a large range, and are not comparable between different datasets, as they depend on the number of clusters found in E . In order to make them more comparable and easier to visualise, I transform the elements of M as follows:

$$M_{ij} \rightarrow \hat{M}_{ij} = \log_2 \left(\frac{1}{N_E} M_{ij} + 1 \right)$$

where N_E is the number of clusters found in E .

I tested this method by creating a dataset that samples data from several multivariate normal (MVN) distributions and combines them. The elements sampled from each MVN therefore form natural clusters in the dataset. By changing the covariance of the MVN distribution, the mixing of the datasets is increased, thus making accurate clustering more difficult.

This dataset is clustered, and then split into two datasets based on the results. Some clusters are placed in one dataset (A), some in another (B), and the remaining clusters are divided between the two. In this way, A and B contain data that should form corresponding clusters (the clusters split between A and B), and

data that certainly do not correspond (the clusters that are placed in either A or B). Hence, clustering A and B separately, we can infer a true mapping between clusters.

I then apply the cluster comparison algorithm described above to A and B . By choosing a score threshold to decide whether two clusters correspond or not, we can compare the found cluster mapping to the true mapping, and thus calculate the true positive rate (TPR) and false positive rate (FPR). By using different covariance values and different score thresholds, I was able to measure the method's performance on noisy data, and find the optimal value for a score cutoff.

The testing results are shown in Figure 3.6 and Table 3.3. As expected, choosing very low score thresholds results in high false positive rates, whereas overly stringent thresholds cause true positives to be missed. A threshold of 7 seems to perform well across the covariance values, even when noise is high, with false positive rates at 0 and true positive rates at 1. It should be noted that in order to maintain reproducibility and reliable clustering structure, the testing data is somewhat artificial; however, it does indicate that the method performs well, and gives a guideline for choosing a score threshold for correspondence. Visual inspection of results from real data also suggests that this method performs well (see Figure 5.11 for an example).

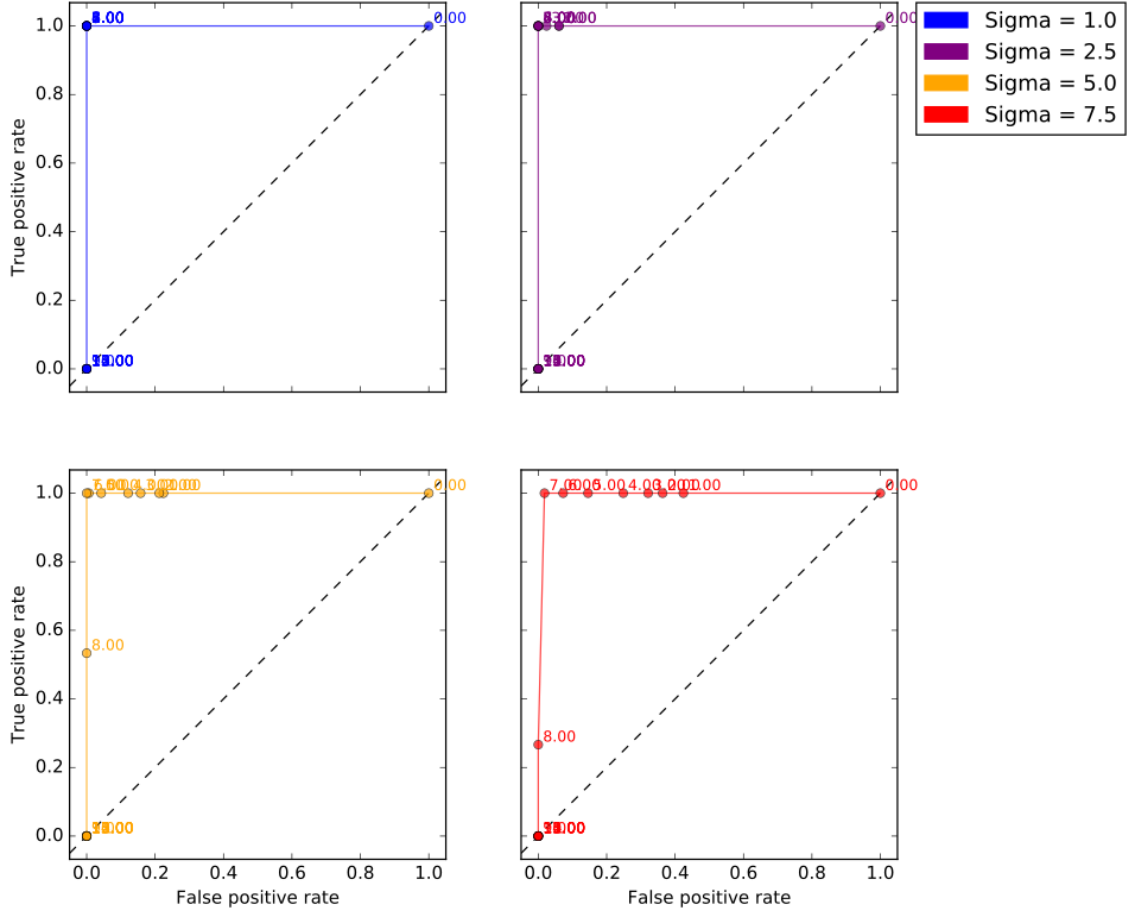


Figure 3.6: Receiver operating characteristic (ROC) plots showing the performance of the cluster comparison algorithm with different covariance values and different score thresholds. Sigma represents the value used to construct the covariance matrix for the MVN distributions. As the covariance increases and clustering becomes more noisy, the false positive rate (FPR) increases; however, using a score threshold of 7.0 produces optimal true positive rates (TPRs) in every case, and low FPRs.

Covariance	Threshold	TPR	FPR
1.0	1	1.00	0.00
1.0	4	1.00	0.00
1.0	7	1.00	0.00
1.0	10	0.00	0.00
1.0	13	0.00	0.00
2.5	1	1.00	0.06
2.5	4	1.00	0.00
2.5	7	1.00	0.00
2.5	10	0.00	0.00
2.5	13	0.00	0.00
5.0	1	1.00	0.22
5.0	4	1.00	0.12
5.0	7	1.00	0.00
5.0	10	0.00	0.00
5.0	13	0.00	0.00
7.5	1	1.00	0.42
7.5	4	1.00	0.25
7.5	7	1.00	0.02
7.5	10	0.00	0.00
7.5	13	0.00	0.00

Table 3.3: A representative subset of the results from testing the cluster comparison algorithm (full results can be found in Online Resources). This confirms the observations from Figure 3.6 that 7.0 is a good choice of score threshold, as it produces optimal TPRs with minimal FPRs (usually zero or close to zero).

3.10 Expression Correlation Distributions

In order to assess the effect of retrotransposon-containing transcripts on protein-coding genes, I paired StringTie transcripts with Ensembl protein-coding transcripts. Pairings were based on the results from gffcompare, a program included with StringTie that matches reconstructed transcripts to reference transcripts, in one of several categories (see Chapter 4).

Each pair could be further categorised based on whether the StringTie transcript contained retrotransposon (RT) sequence, and if so, whether the Ensembl transcript also contained matching RT sequence. For each category, I obtained kallisto expression estimates for the StringTie and Ensembl transcripts for each sample individually. I then filtered pairs based on expression estimates (both with TPM >1), and the number of samples the pair was expressed in (>5). I calculated the Spearman’s rho correlation coefficient [233] for that pair across the relevant samples. This produced a set of correlation coefficients for each category. I then plotted the distribution of these values. I used an Anderson-Darling test to test whether the difference between the distributions was statistically significant.

3.11 Venn Diagrams and Statistical Analysis

Venn diagrams were created using the matplotlib-venn package [234]. Visual inspection of a Venn diagram can suggest a bias towards sharing or non-sharing between categorised sets of interest (e.g., retrotransposons expressed in a given cell type). To assess these possibilities statistically, I used the following approach. First, I generated sets for each category by drawing randomly from the set of

all possible members, and counted the number of overlaps between categories. I repeated this to produce a null distribution, representing the degree of overlap between categories expected by chance. I visualised the actual observed overlaps compared to the randomly generated data, and then calculated p values representing the likelihood that the observed value was drawn from the null distribution. To do this, I first checked whether the randomly generated data could be modelled by a normal distribution, using quantile-quantile plots, the Shapiro-Wilk test for normality [235], and the D’Agostino-Pearson test for normality [236]. If these checks showed that a normal distribution was appropriate, I calculated the parameters for the normal distribution that would model the random data. Using these parameters, I applied the cumulative density function (CDF) to the observed values. This returns a value representing how likely it is to observe a value at least as extreme as the input from the normal distribution. This can be used as a p -value.

I also compared Venn diagrams where one represents a subset of the data represented in the other. In particular, I wanted to quantify whether the subset was split in the same proportions as the larger set. Suppose D is the larger dataset, which has been split into categories as shown in a Venn diagram, and E is a subset of D . We can use the proportion of D in each category to calculate the expected number from E in each category. This represents the null hypothesis that E follows the same distribution as D . We then categorise E and use a chi-square test to compare the observed and expected values. A small p value suggests that we should reject the null hypothesis, and that E follows a different distribution to D .

3.12 Retrocopy Transcripts

To identify retrocopies that are transcribed and the transcripts they belong to, I used a similar method to that used to identify transcribed retrotransposons. As with retrotransposons, the retrocopy annotations I used have internal structure. The gaps between these blocks can include non-retrocopy regions, such as retrotransposons. Again, naive intersections between transcripts and retrocopies could produce false-positives. I used the following method to accurately identify true retrocopy transcripts (steps 1 and 2 are identical to the above method):

1. Apply `bedtools intersect` to a set of exons and a set of individual retrocopy blocks
2. Reassemble exons into transcripts and reassemble retrocopy blocks into full elements, preserving the overlap relationships
3. Filter overlaps based on proportion of total exon/block length covered by the intersection, for both transcript and retrocopy

3.13 Retrocopy Conservation

In order to assess the degree to which retrocopies have retained sequence identity with their parent transcripts, I carried out an alignment between each retrocopy and its parent. This was done using the `matcher` software, part of the EMBOSS suite [237], wrapped in a custom script to achieve parallelisation and to manage the inputs and outputs efficiently. This produces a file for each pair containing information about the best alignment between the two.

Each alignment has two important properties for assessing how good it is:

- Length: the length, in nucleotides, of the match
- Identity: the proportion of nucleotides in the matching regions which are the same in both the retrocopy and parent.

Separately, these can both be misleading measures of alignment. An alignment may be the full length of the query, but with very low identity. Conversely, an alignment may have 100% identity but represent only a short subsequence of the query. Neither of these represent true alignments between a retrocopy and its parent. I therefore combined these two to produce an alignment score, S :

$$S = \frac{MI}{Q} \quad (3.2)$$

where M is the length of the match, I is the identity, and Q is the length of the query sequence (in this case, the retrocopy). This represents the proportion of nucleotides in the retrocopy that have a match in the parent transcript. This measure is therefore high for long matches with high identity. It is not confounded by gaps resulting from lost introns, but will penalise acquired mutations.

3.14 Retrocopies in CAST

In order to find BL6 retrocopies conserved in CAST, I extracted the sequences for all BL6 annotated retrocopies (query sequences) and searched for matches in the CAST genome (subject) using `blastn`. This produced 6,645,353 hits across 18,139 BL6 retrocopies. However, many of these hits were due to the alignment of a short subsequence of a query, and do not represent conservation. Many of the hits were also interchromosomal: the chromosome of the query sequence did not

match that of the subject. Again, these hits do not represent conservation of the query. I therefore removed all interchromosomal hits, and filtered the remaining hits on the length of the match as a proportion of the length of the query, requiring this to be > 0.8 . These filters reduced the number of hits to 27,875 across 10,694 retrocopies.

Figure 3.7 shows a visualisation of these hits. A large number of hits show significant shifts in the relative position in the chromosome of the query versus the match, which suggests either a false positive or a structural variation that has affected the retrocopy. I therefore removed hits where the relative position of the query did not closely match the relative position of the match. The threshold was chosen by considering Figure 3.8. There is a clear enrichment along the diagonal representing true-positive hits. Cutoffs were chosen to include points in this region, while excluding the others. This further reduced the number of hits to 22,312 across 10,493 retrocopies.

A more sophisticated version of this analysis could use a list of known structural variants (SVs) between BL6 and CAST to identify those retrocopies that had been affected by an SV, and thus include them in downstream analysis. This would expand the range of retrocopies, increasing the power and reliability of downstream analysis. One could also compare those that had been affected by an SV and those that had not. Due to the complexity of the SV catalogue I was unable to include such an analysis here.

While there are 18,456 retrocopies in the BL6 annotation I used, these originate from just 3,860 parent transcripts, notwithstanding those retrocopies that do not have a definitively labelled parent. This implies that there will be distinct retrocopies from the same parent with high sequence identity, which would there-

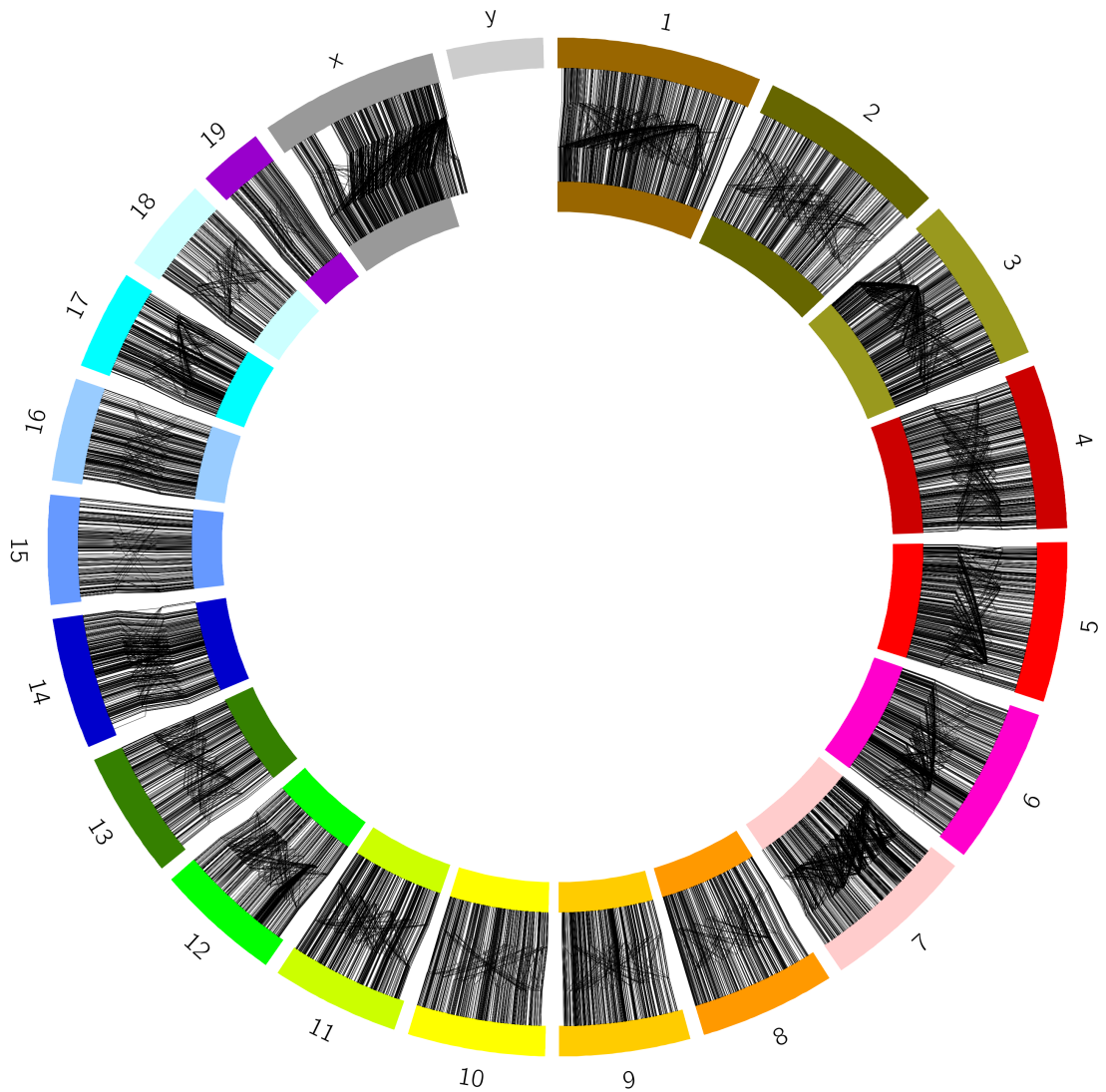


Figure 3.7: BL6 retrocopy matches in CAST. The outer ring represents the BL6 genome, while the inner ring is CAST. Lines link BL6 retrocopies to their matches in CAST, after filtering on length of match and removing interchromosomal hits. A large number show significant shifts in relative position on the chromosome. Figure created using the Circos software [238].

fore have sequence matches at the same loci in the CAST genome, erroneously increasing the number of hits. To remove these false hits, I merged the match regions based on location, so overlapping hits are merged into a single region. This

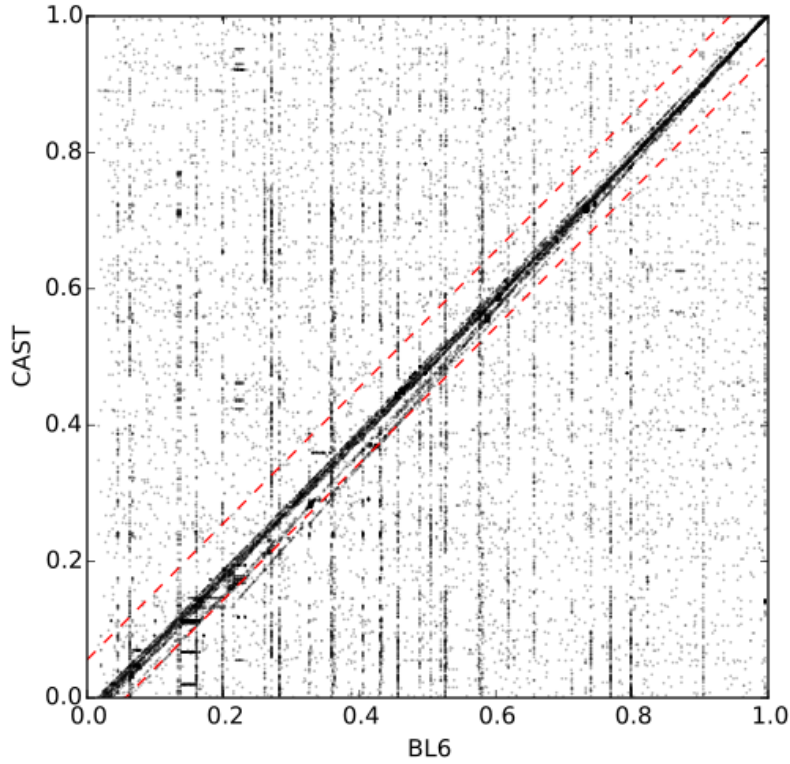


Figure 3.8: The relative positions on chromosomes for retrocopies in BL6 and their matches in CAST, after filtering for length and removing interchromosomal matches. Red dashed lines indicate the cutoff for keeping a match.

resulted in a reduced list of 14,788 regions, but with some regions having multiple matching BL6 retrocopies. I split these regions based on whether the matching retrocopies shared a single parent (11,475), did not share a single parent (485), or had an unknown parent (2,558).

At this stage, I had a list of 11,475 regions representing BL6 retrocopies conserved in CAST. However, 12.1% of these matched multiple BL6 retrocopies, and so I could not reliably assign these a single retrocopy. However, they each had an agreed parent transcript in BL6, with 2,666 parent transcripts accounting for the

11,475 regions.

It is important to note that this is not an exhaustive search for retrocopies in CAST, but only a search for those conserved between BL6 and CAST that have retained their relative position in the chromosome. While I have endeavoured to remove as many false positives as possible, some may still remain; in particular, retrocopies may match their parent transcript. Intron loss and acquired mutations will have made it less likely for retrocopies to match their parents, and removal of interchromosomal matches will have also reduced the number of retrocopy/parent matches. The process used to remove false positives is likely to be overly strict, and ignores the possible effects of SVs, as discussed above.

3.15 Computing Environments and Online Resources

I carried out the majority of computing work for this thesis on the Ubuntu 14.04 operating system, and the remainder on recent versions of macOS. The majority of the custom code I developed is written in Python 2.7, with additional scripts written using the R programming language, the Bash shell, Perl, and other command line tools included with Ubuntu; in particular `sed`, `awk`, `xargs`, and `parallel` [239]. Particular note should be given to the superb `matplotlib` [240], `numpy` [241], `scipy` [242], and `scikit-learn` [243] Python packages, which I used extensively. This thesis was typeset using \LaTeX .

Original code, processed data, and additional images can be found in the Online Resources at <https://github.com/jmg1297/thesis>. For larger datasets (e.g.,

raw reads, whole transcriptomes), contact the Ferguson-Smith lab directly.

Chapter 4

Mouse Lymphocyte

Transcriptomes

I ran the analysis pipeline described in Methods on the BLUEPRINT BL6 and CAST samples, resulting in alignments and transcriptomes for B and T lymphocytes in males and females from two diverged strains. This represents a useful resource for future analysis, particularly in reference to retrotransposons and other repetitive content.

4.1 Alignments

I aligned the raw reads for each BLUEPRINT BL6 and CAST sample and removed reads mapping to ribosomal RNA (rRNA). I retained uniquely mapping reads and multimapping reads with fewer than 50 matches for downstream analysis. After alignment, the reads are divided into the four categories shown in Figures 4.1 and 4.2:

- Uniquely mapped: reads that mapped to a single location in the reference genome
- Multimapped: reads that mapped to multiple locations in the reference genome
- rRNA: reads that mapped to ribosomal RNA regions and therefore needed to be removed before further analysis (see Methods)
- Unmapped: reads that could not be mapped to any location in the genome, or that mapped to too many locations in the genome

There is significant variation in the amount of rRNA present in each sample, both in terms of raw read numbers and percentage of total reads. In some cases, this significantly reduces the number of reads available for downstream analysis. The worst example of this is a reduction from 80 million reads to under 20 million. While this represents a significant reduction in coverage, using these in combination with the other samples should still produce reliable results, and so these samples were retained.

Figures 4.3 and 4.4 show the distributions of alignment score (AS), representing the quality of the alignment based on length and number of mismatches, and number of hits (NH), the number of loci each read is mapped to, for the BL6 and CAST samples. AS scores greater than or equal to 200 are ideal, representing a full-length (or near full-length) paired-end match, and NH would be 1 for a unique alignment. Both show all of the samples following similar distributions, with high AS values (around 200) and low NH values. Some of the CAST samples show weaker peaks around 200 for AS values, and tend to have fewer multimapping

reads. This could be a result of the lower quality of the CAST genome compared to the BL6 genome.

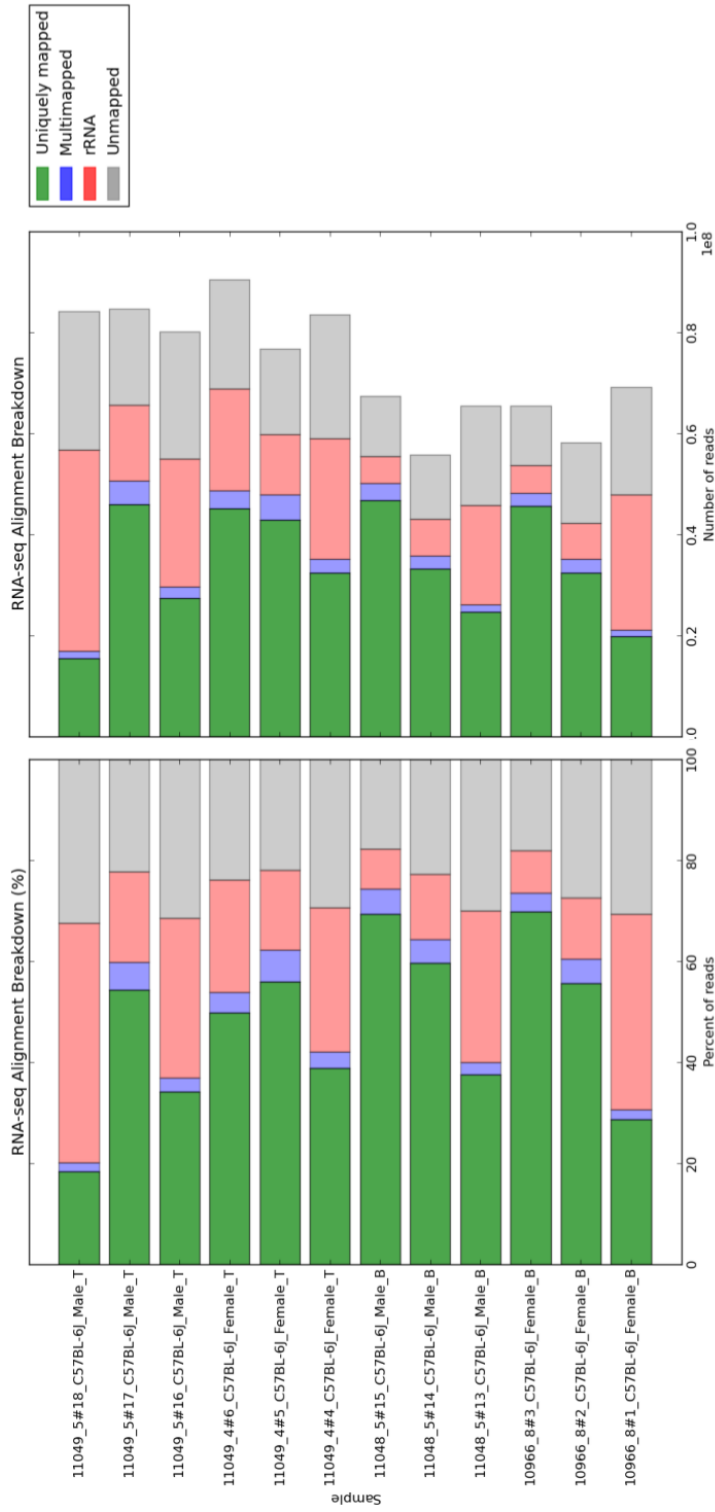


Figure 4.1: Summary of alignments for the BLUEPRINT BL6 samples. The left-hand figure shows proportions of reads falling into each category, while the right shows the number of reads. In some cases there is significant reduction in coverage due to removal of rRNA and unmapped reads.

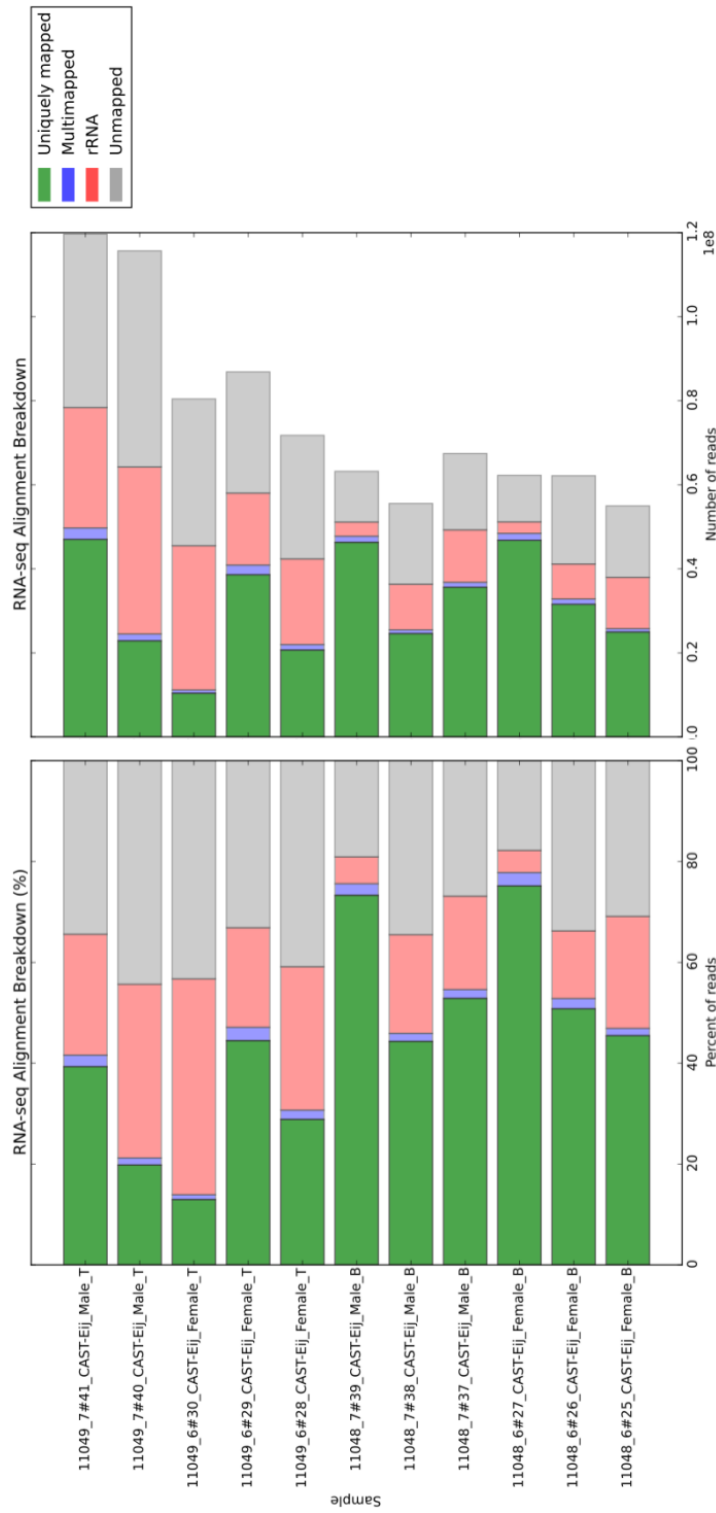


Figure 4.2: Summary of alignments for the BLUEPRINT CAST samples. The left-hand figure shows proportions of reads falling into each category, while the right shows the number of reads. As in the BL6 samples, all samples have reduced coverage due to rRNA and unmapped reads. This suggests that the method used to identify rRNA regions in CAST was effective.

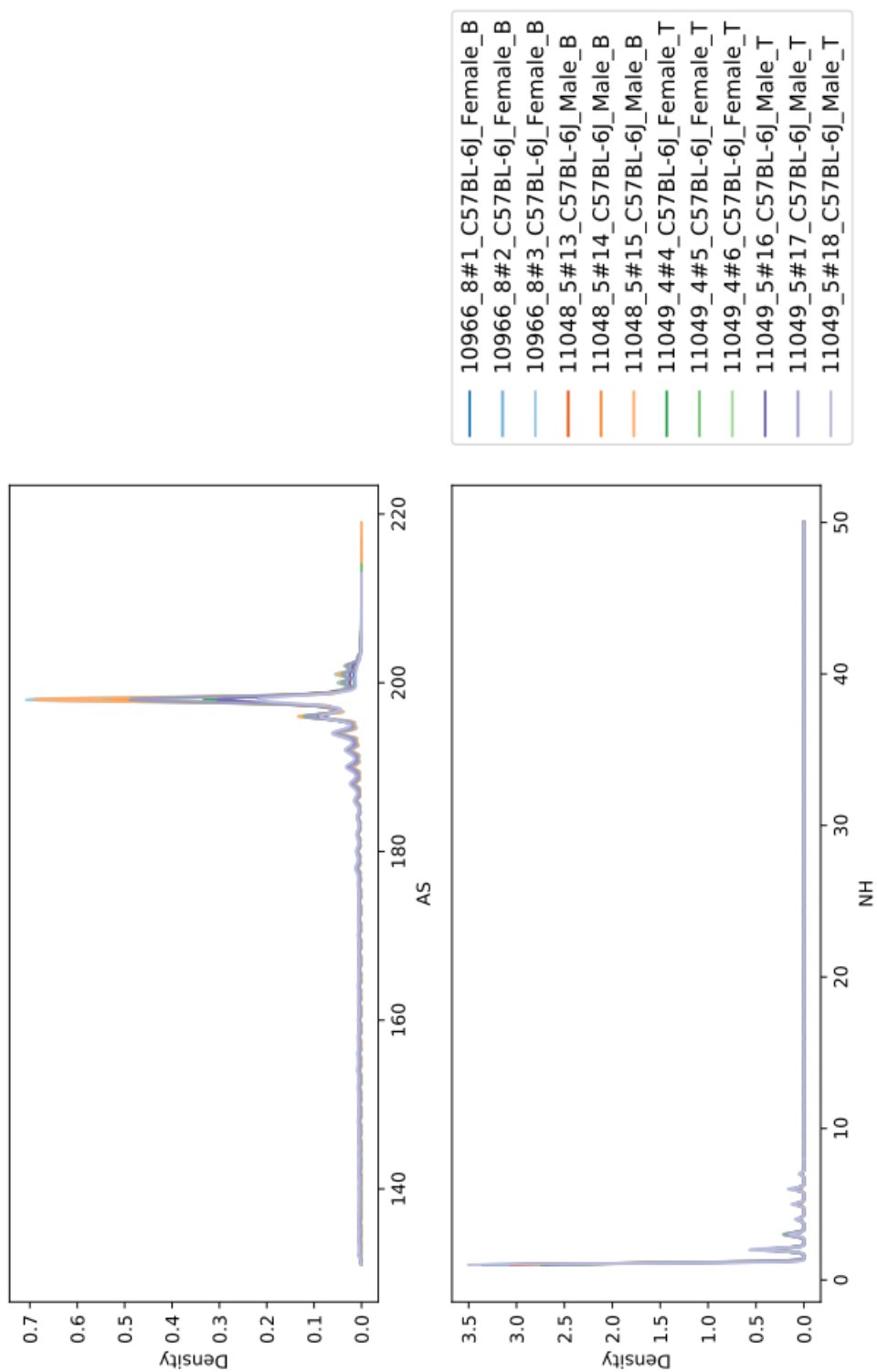


Figure 4.3: Distribution of alignment score (AS) and number of hits (NH) for the BL6 samples. In all samples, the majority of alignments are high scoring around 200, and the majority of reads map uniquely, indicating that the alignments are good quality.

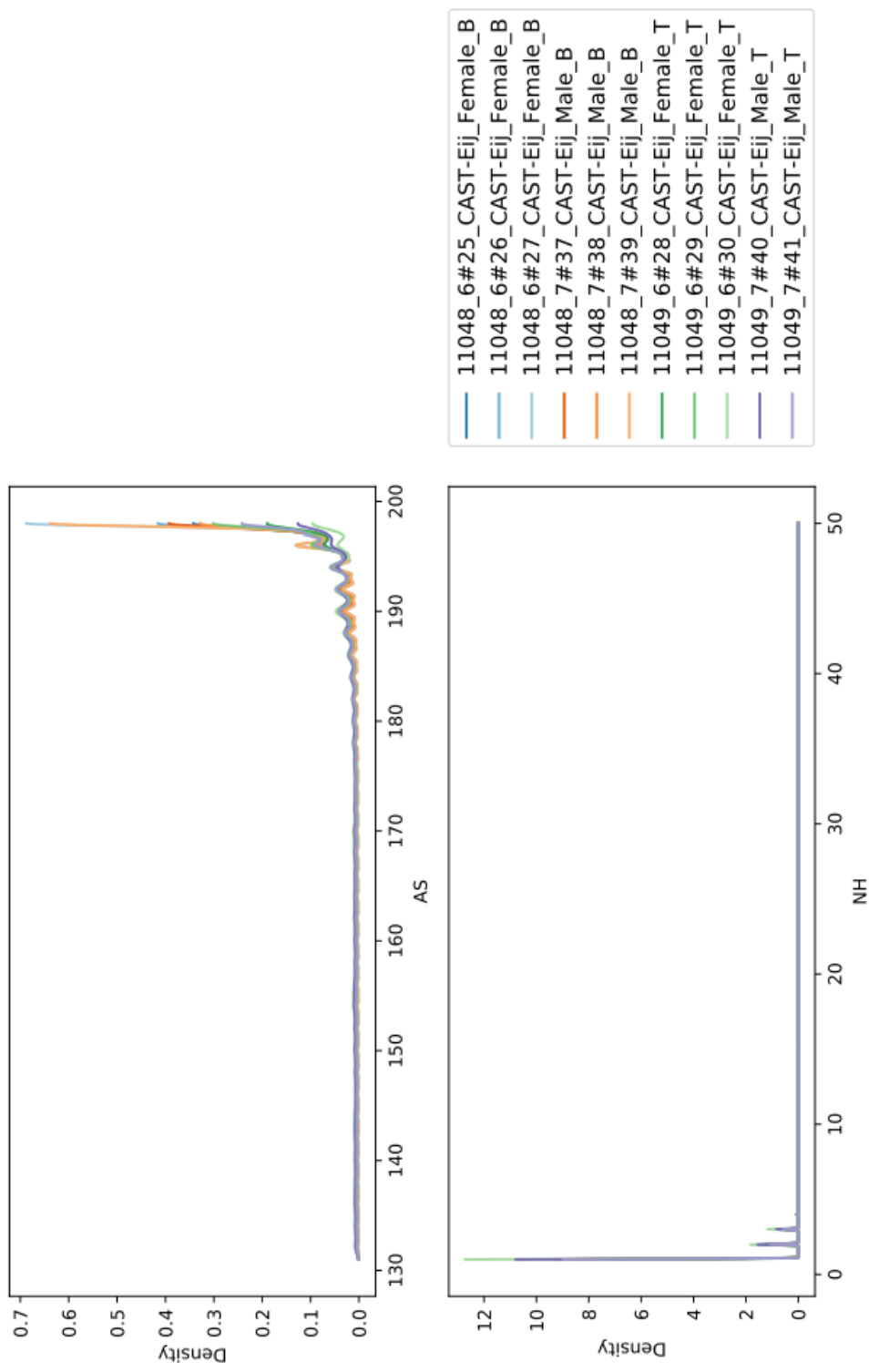


Figure 4.4: Distribution of alignment score (AS) and number of hits (NH) for the CAST samples. The distributions of AS are not as good as for the BL6 samples, with no scores above 200 for any samples. This may be due to the lower quality of the CAST reference genome compared to the BL6 reference. However, scores are still high, with the majority of alignments scoring near 200. As in BL6, the majority of the CAST alignments are unique. Overall, the results still indicate a good alignment for the CAST samples.

4.2 Reconstructed Transcriptomes

I produced a transcriptome for each BLUEPRINT BL6 and CAST sample, and merged transcriptomes based on phenotype combinations. The results are summarised in Figures 4.6, 4.7, 4.8, and 4.9. I used these summary figures to check for consistency in the features found in each reconstructed transcriptome. Each figure contains six subfigures:

- Top left, transcripts per gene: the number of distinct transcripts for each reconstructed gene
- Top middle, exons per transcript: the number of exons for each reconstructed transcript
- Top right, transcripts per chromosome: the number of transcripts found in each chromosome
- Bottom left, total genes: the total number of reconstructed genes in each sample
- Bottom middle, total transcripts: the total number of reconstructed transcripts in each sample
- Bottom right, total exons: the total number of reconstructed exons in each sample

The feature numbers are consistent across individual samples and between merged transcriptomes. As expected, merged transcriptomes including more samples contain more features. The female samples and female-only merges contain a small number of transcripts on the Y chromosome, although fewer than those

from the male samples and male-only merges. Figure 4.5 shows the distribution of the number of hits per read (i.e., the number of locations to which the read maps) for male and female samples. The reads mapping to the Y chromosome in the female samples are more likely to have multiple hits, compared to those in the male samples. This suggests that many of the erroneous Y mappings in the female samples are in fact due to the fact that the pipeline retains multimapped reads. The Y chromosome reads and transcripts in the female samples should be ignored in downstream analysis.

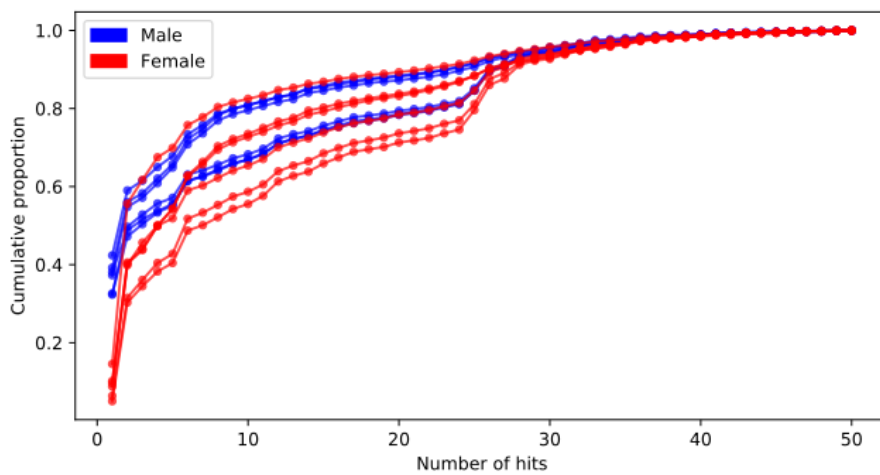


Figure 4.5: The cumulative distribution of NH values in the male and female BL6 samples. While the higher NH values do not clearly differ between male and female values, the female samples have fewer uniquely mapping reads on the Y chromosome and more with at least 2 mappings.

For the BL6 samples, I compared the merged transcriptomes to the Ensembl reference transcriptome to ascertain which of the reconstructed transcripts were novel, summarised in Figure 4.10. The results are consistent across the merges. Approximately 60% of transcripts in each merge matched an Ensembl transcript, either completely or with at least one shared splice junction. Of those remaining,

the majority are antisense to a known transcript, or a novel intergenic transcript. In the transcriptome merging all samples, the transcripts matching a reference cover 11,933 of the possible 115,220 Ensembl transcripts.

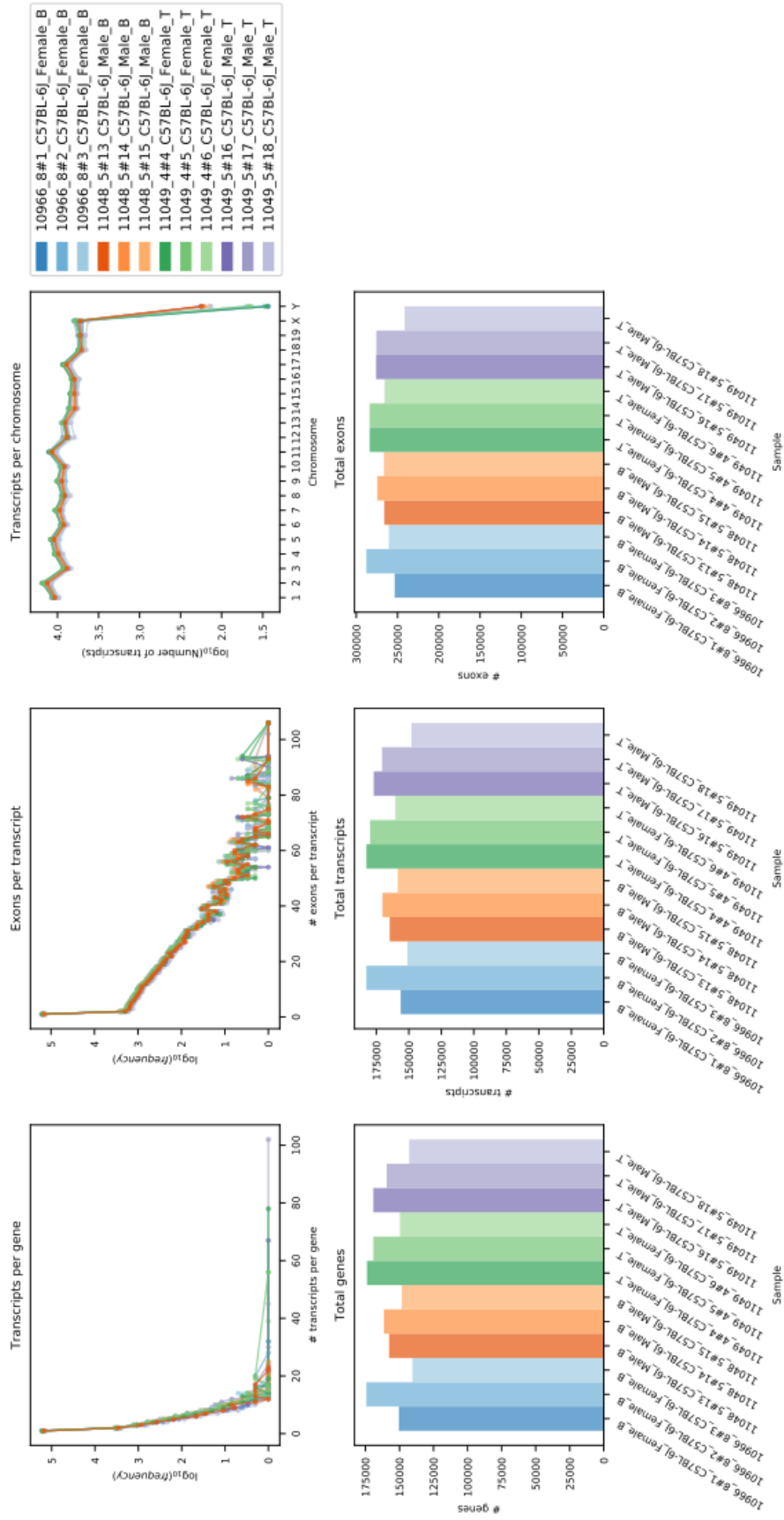


Figure 4.6: Transcriptome summaries for each BL6 sample. All samples show consistent features in the reconstructed transcriptomes.

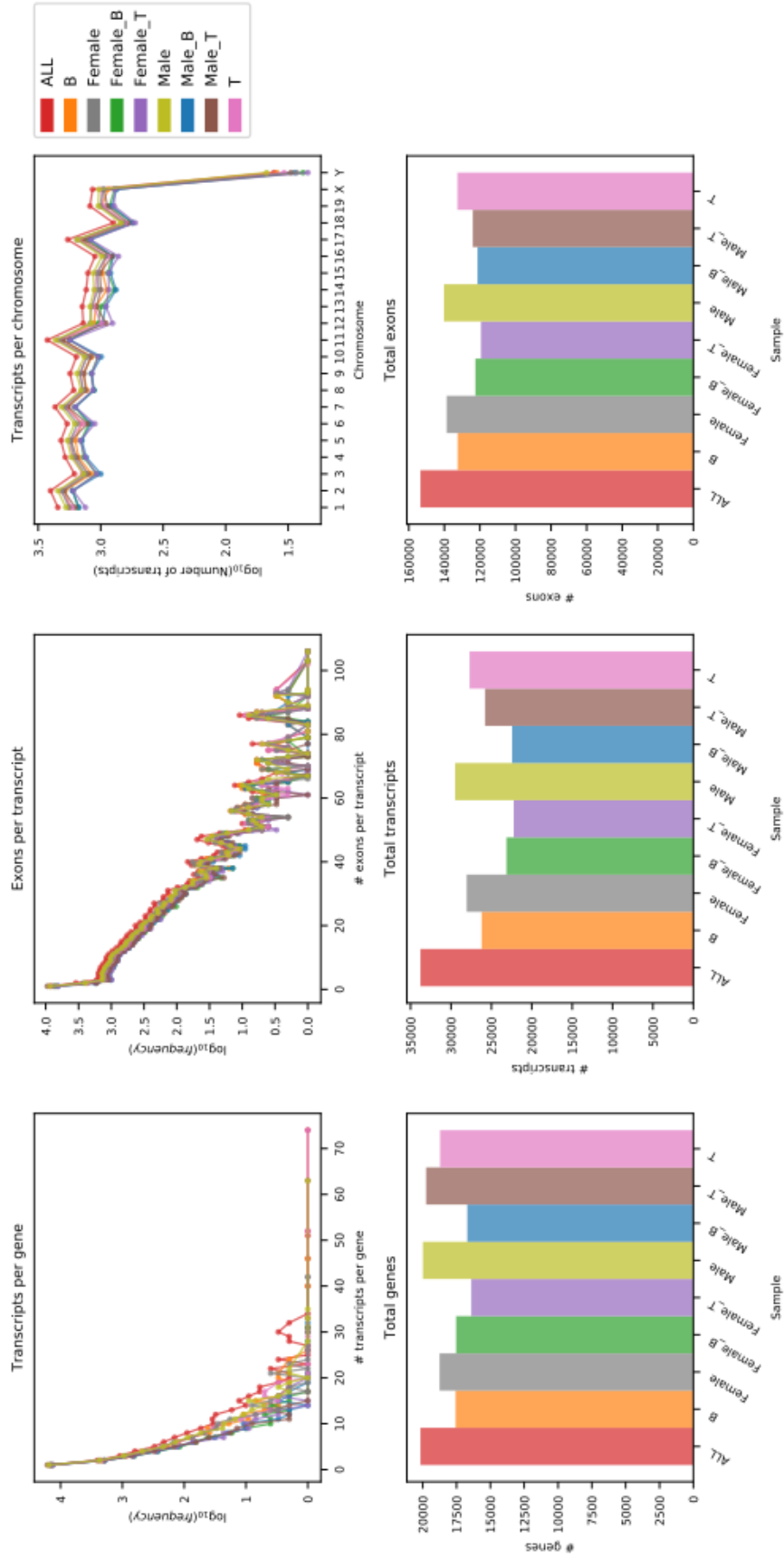


Figure 4.7: Transcriptome summaries for merged BL6 transcriptomes. As expected, the transcriptome merging all samples (labelled ALL) has the highest number of features in all categories. All of the visualised features are consistent across the merges.

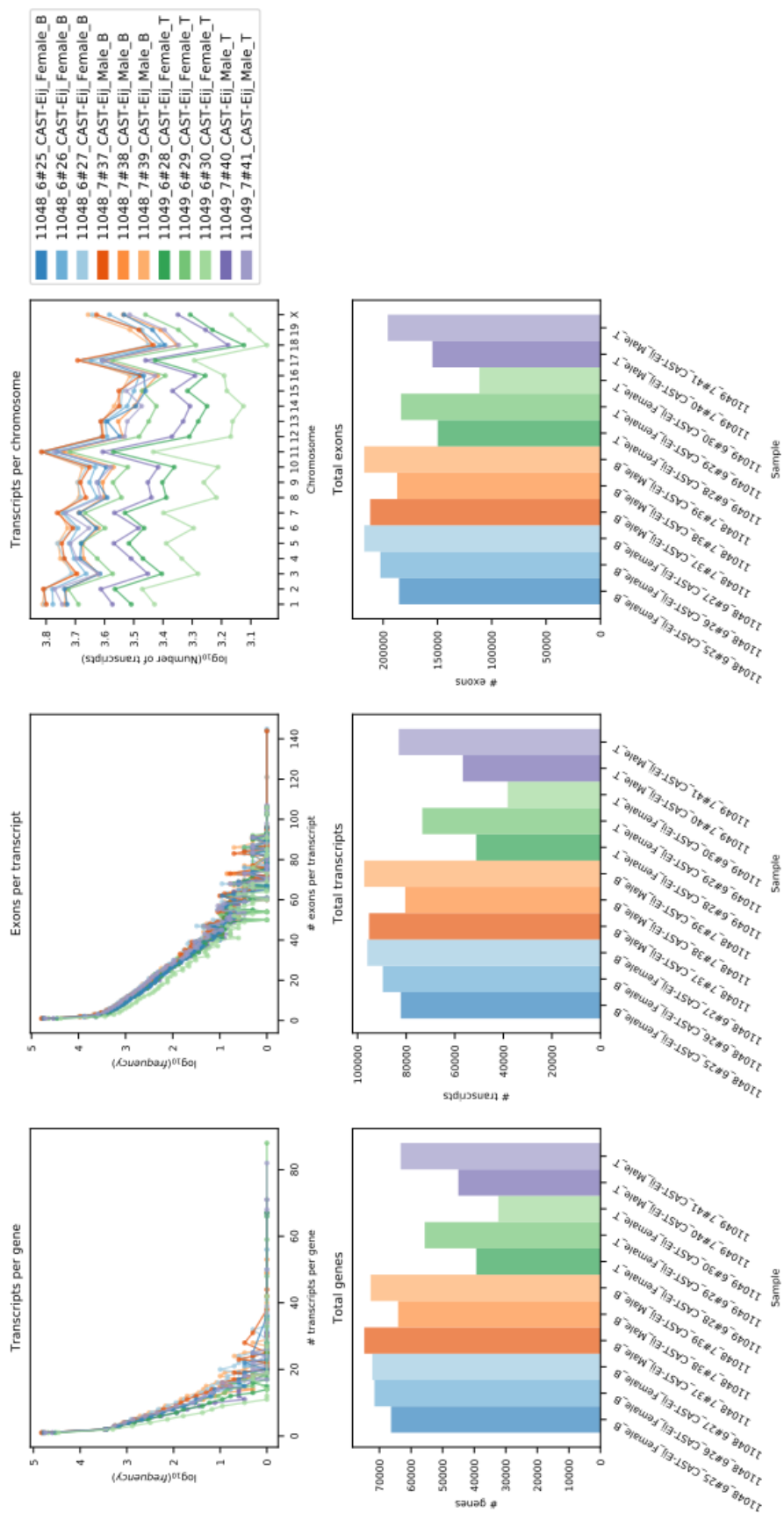


Figure 4.8: Transcriptome summaries for each CAST sample. There is more variation in the samples here compared to the BL6 samples, possibly due to the lower quality of the CAST reference genome. This is particularly true of the “transcripts per chromosome” figure (top right). However, every sample follows the same pattern across the chromosomes.

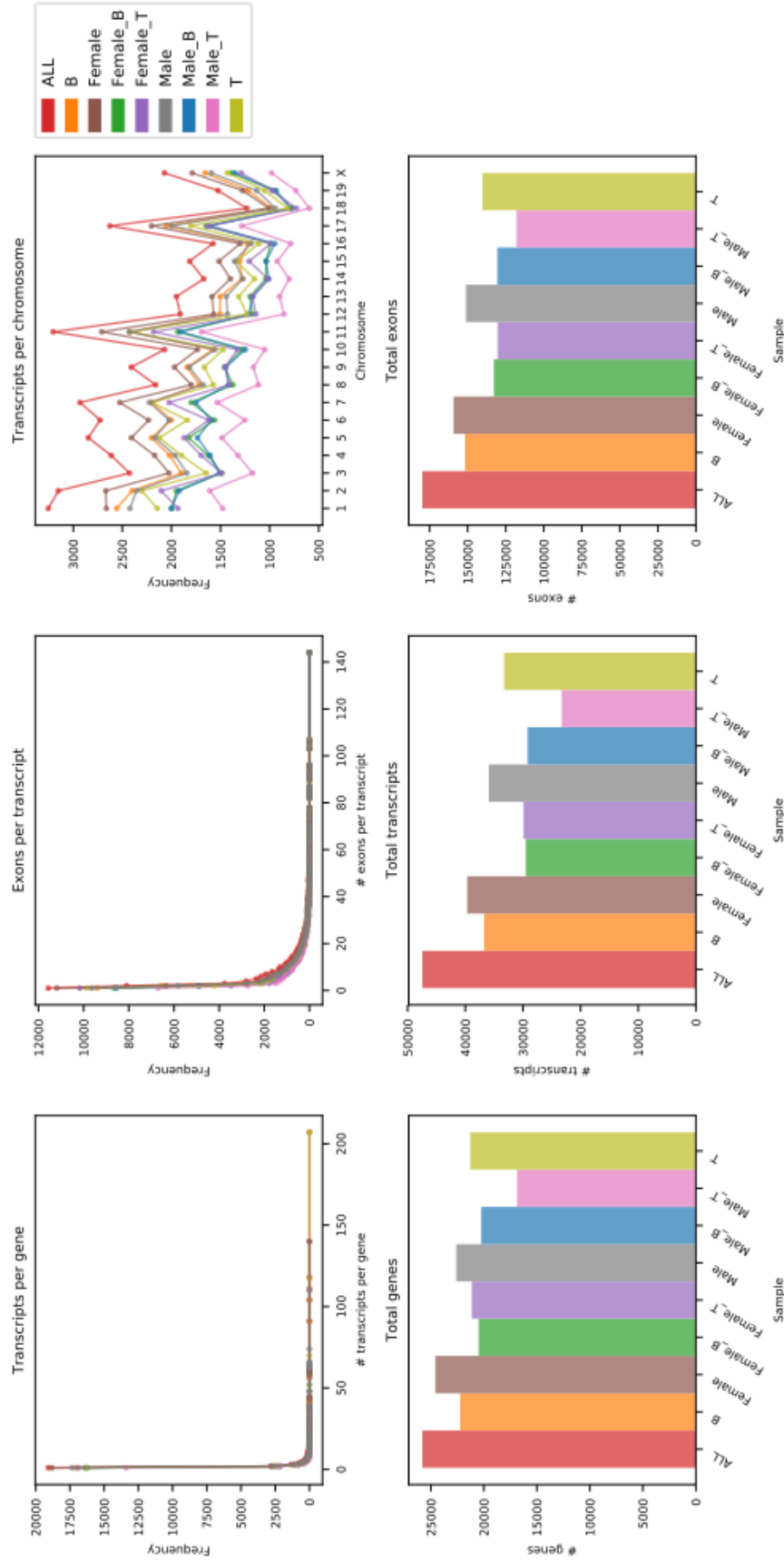


Figure 4.9: Transcriptome summaries for merged CAST transcriptomes. As in the per-sample CAST summaries, there is more variation than in the equivalent BL6 analysis; this is to be expected given the variation between the individual samples. A notable difference from both the BL6 merges and the individual CAST samples is in the “transcripts per gene” and “exons per transcript” figures (top left, top middle). These show a much stronger bias towards single-transcript genes and single-exon transcripts. This is not the case in the individual CAST samples, suggesting that the isoforms and exon junctions found in the individual samples are not consistent, and so cannot reliably be called in the merged transcriptomes. This may be due to less reliable mapping to the lower-quality CAST genome.

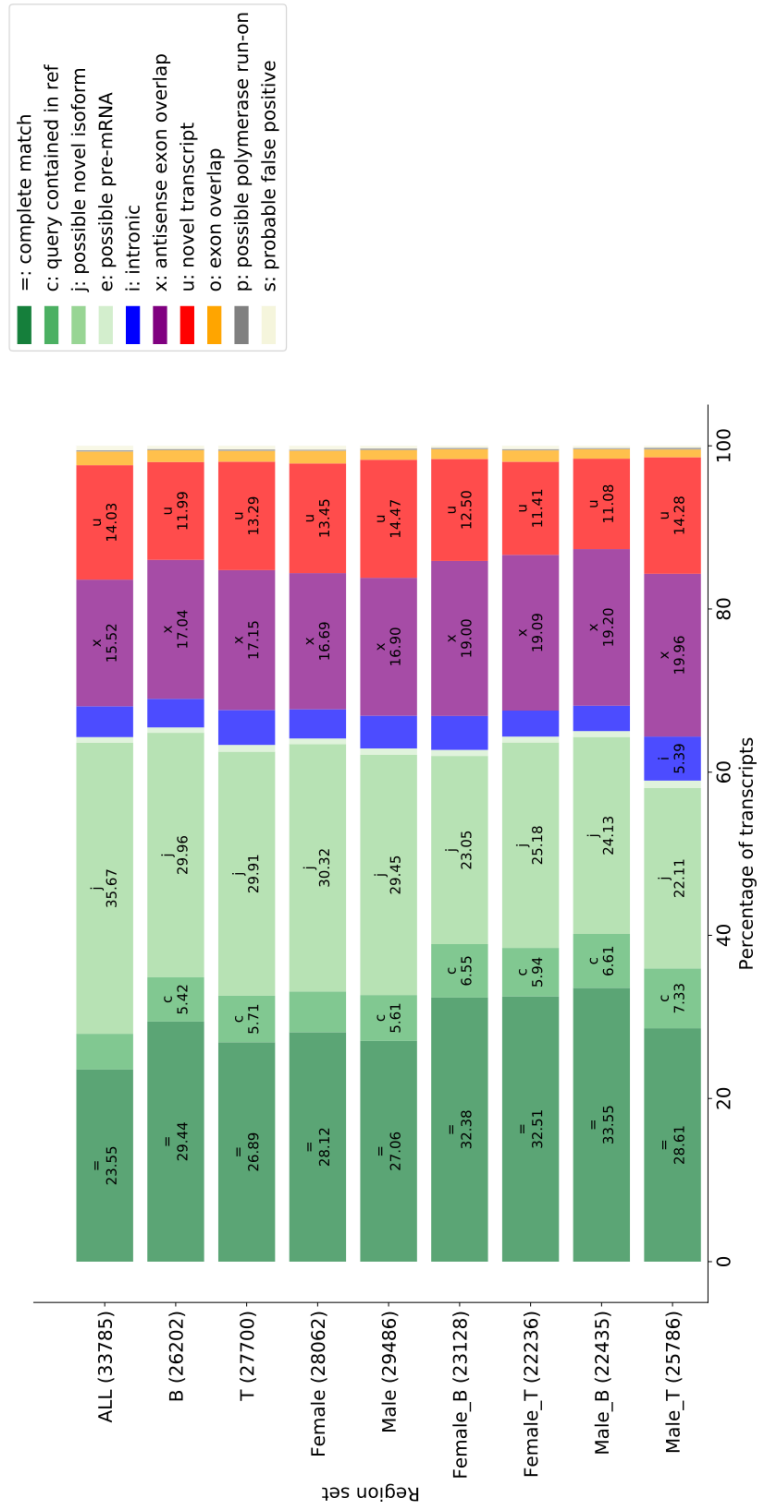


Figure 4.10: Breakdown of comparison between BL6 merged reconstructed transcriptomes and the Ensembl reference annotation.

Chapter 5

Retrotransposon Transcription in Lymphocytes

As discussed in the Introduction, many of the studies on retrotransposon transcription have focused on pluripotent and embryonic cells, with relatively few studies on transcription of retrotransposons in somatic cells. Those that have focused on somatic cells have found evidence of retrotransposon activity in multiple cell types, and have identified a close link between retrotransposons and long non-coding RNAs (lncRNAs) [127–129]. The high quality BLUEPRINT RNA-seq datasets provide an ideal opportunity to compare retrotransposon transcription in two somatic cell types in detail. Using the reconstructed transcriptomes described in Chapter 4, I aim to:

- Quantify the retrotransposon content in the BLUEPRINT BL6 transcriptomes
- Test the hypothesis that retrotransposon transcripts affect gene expression

levels

- Compare the retrotransposon content of B and T lymphocytes

For these analyses I used only BL6 samples, as high quality retrotransposon and gene annotations are readily available for this reference genome. To create these annotations in CAST with comparable quality was beyond the scope of this project; however, obtaining these annotations and running a similar analysis in CAST should be included in future work.

5.1 Quantifying Retrotransposon Transcription

In order to quantify retrotransposon transcription in B and T lymphocytes, I compared the merged reconstructed transcriptomes to the RepeatMasker retrotransposon annotation and used novel software tools to visualise the results, as described in Methods.

The results are summarised in Table 5.1. I found between 8,600 and 16,000 transcripts with exons overlapping retrotransposons, representing 40-45% of the transcripts in the reconstructed transcriptomes. However, for the vast majority of these transcripts there is only a small amount of overlap between their exons and retrotransposons: only 3-5% have an overlap of more than 50% with retrotransposons. Of these, about half have an overlap of more than 90% (see Table 5.1 and Figure 5.1).

Including all retrotransposon-containing transcripts, agglomerative clustering reveals little structure in the data, dividing the transcripts into three clusters: two with relatively high retrotransposon content, and one large and noisy cluster

Transcript Set	Total Transcripts	Transcripts with RT		> 50% RT		> 90% RT	
		Number	%	Number	%	Number	%
ALL	33,785	16,225	48.02	1,757	5.20	868	2.57
B	26,202	11,207	42.77	948	3.62	404	1.54
T	27,700	12,382	44.70	1,374	4.96	698	2.52
Females	28,062	12,806	45.63	1,244	4.43	609	2.17
Males	29,486	12,934	43.86	1,517	5.14	770	2.61
Female B	23,128	9,300	40.21	829	3.58	333	1.44
Female T	22,236	9,317	41.90	813	3.66	445	2.00
Male B	22,435	8,645	38.53	657	2.93	283	1.26
Male T	25,786	10,506	40.74	1,387	5.38	687	2.66

Table 5.1: A summary of the BL6 reconstructed transcriptomes and the retrotransposon (RT) content of each. A very small percentage of transcripts overlapping retrotransposons contain more than 50% retrotransposon content, and about half of these contain more than 90% retrotransposon content.

with low to intermediate retrotransposon content. The retrotransposon elements overlapping transcripts are dominated by SINEs, which make up approximately 75% of all of the retrotransposon elements overlapping transcripts. LINEs and LTRs make up approximately 10% and 15% (Table 5.2). These represent an enrichment in SINEs and a depletion in LINEs and LTRs, compared to their contributions to all retrotransposons (Figure 5.4).

However, when transcripts with less than 50% retrotransposon content are removed, the same clustering algorithm shows several distinct clusters based on the type of retrotransposon that the transcripts overlap (Figure 5.2). LINEs and LTRs dominate in terms of transcript sequence content and number of transcripts (Table 5.2 and Figure 5.2). SINEs are no longer overrepresented, and the proportion of LTRs exceeds their proportion of all retrotransposons (Figure 5.4). There is

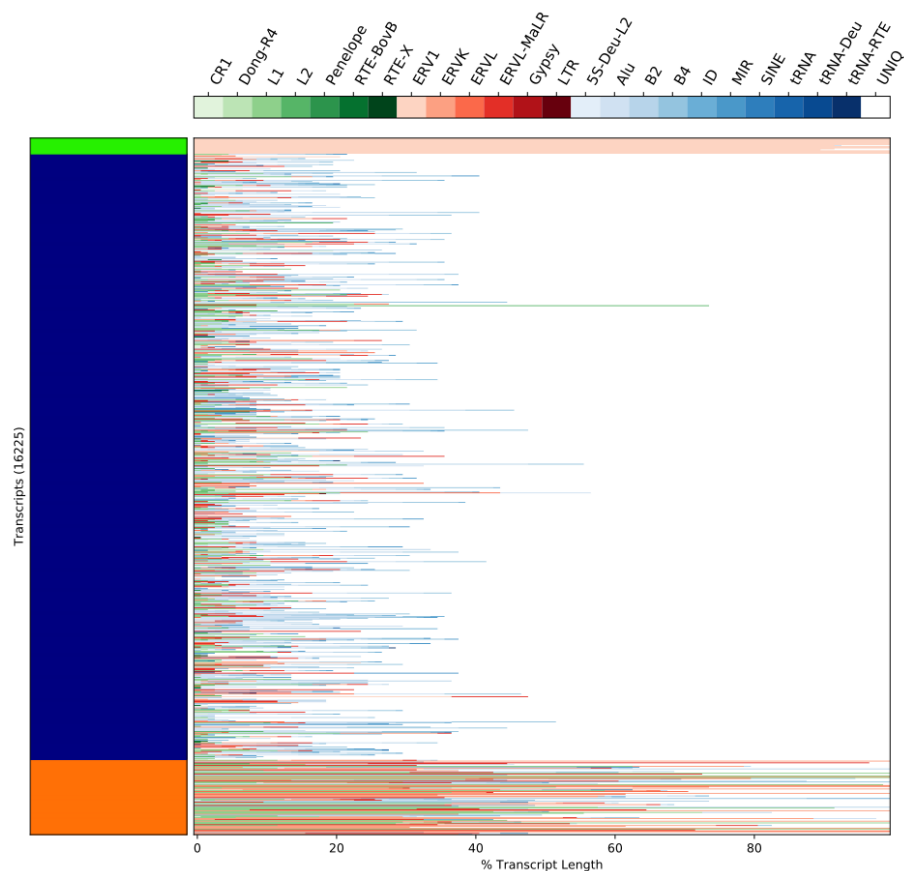


Figure 5.1: A heatmap showing the retrotransposon content of the reconstructed transcriptomes, merged across all BL6 samples. Each row represents one reconstructed transcript, and the bars in that row represent the total proportion of retrotransposon sequence in the exons of that transcript. The colours represent the different classes of retrotransposon, according to the legend at the top of the figure. Greens represent LINES; reds represent LTRs; blues represent SINEs; and white for non-RT sequence. Different shades of each colour represent specific sub-classes. The blocks on the left of the figure show the results of agglomerative clustering based on the total proportion of each type of retrotransposon. The methods used to produce these figures are described in detail in Methods. In this figure, clustering reveals relatively little structure: there is one small cluster of transcripts with high ERV1 content; a larger cluster with high RT content, but a mix of classes; and the large cluster with relatively low RT content.

particular enrichment for ERV1, ERVK, ERVL-MaLR, and L1 retrotransposons. These patterns are even more pronounced when selecting transcripts with more

than 90% retrotransposon content (Figure 5.3).

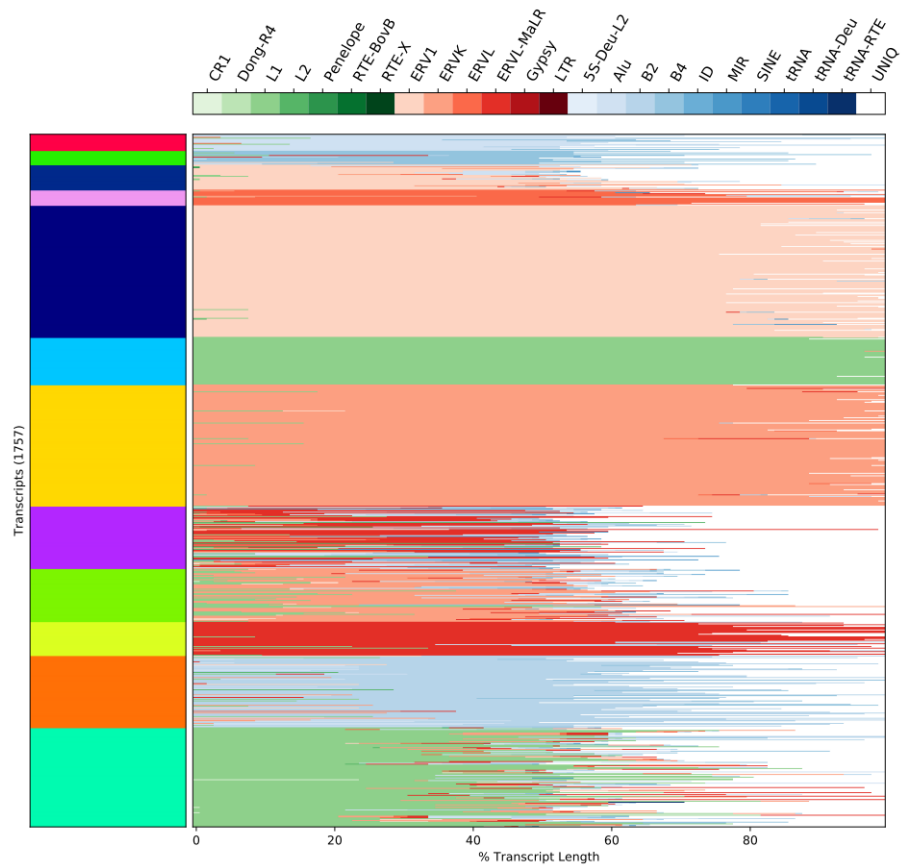


Figure 5.2: Retrotransposon content of transcripts containing >50% RT sequence.

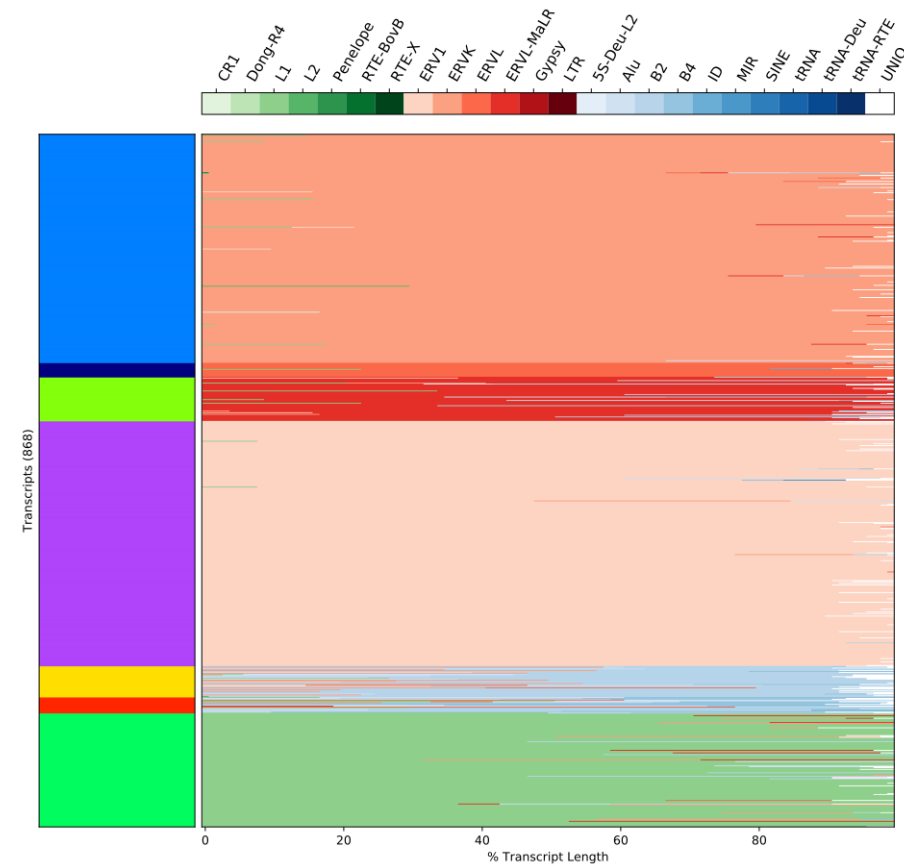


Figure 5.3: Retrotransposon content of transcripts containing >90% RT sequence.

This suggests that retrotransposon-containing transcripts can be divided into two main categories: those with low to intermediate retrotransposon content, and those with high retrotransposon content. Transcripts in the former group tend to have a mixture of different retrotransposons, possibly representing small fragments of retrotransposon or those present in introns and untranslated regions (UTRs). These could be functionally relevant as regulatory regions, or may simply be chance insertions that do not play any significant role.

Transcripts in the latter group can be subdivided into clusters based on the kind of retrotransposon they overlap, and are dominated by LINE and LTR retrotransposons. This suggests the existence of a set of LTR- and LINE-based transcripts. Such transcripts could have significant regulatory potential, based on sequence similarity with RT-derived regulatory regions or RT sequence in other transcripts [244]. Some of these could also be retrotransposition intermediates, as the mouse genome does contain potentially active retrotransposons [55, 245].

Transcript Set	Total RT Elements	LINEs		SINEs		LTRs	
		Number	%	Number	%	Number	%
> 0% RT Content	112,705	11,731	10.41	85,122	75.53	15,852	14.07
> 50% RT Content	5,846	1,105	18.90	2,694	46.08	2,047	35.02
> 90% RT Content	1,089	201	18.46	188	17.26	700	64.28

Table 5.2: The number of individual retrotransposon (RT) elements overlapped by transcripts in the reconstructed transcriptome across all B and T samples. SINEs dominate when including all transcripts with retrotransposon content, but not when filtering on retrotransposon content percentage. In particular, the number of ERV elements increases rapidly, which is reflected in the sequence content (Figures 5.1, 5.2, and 5.3).

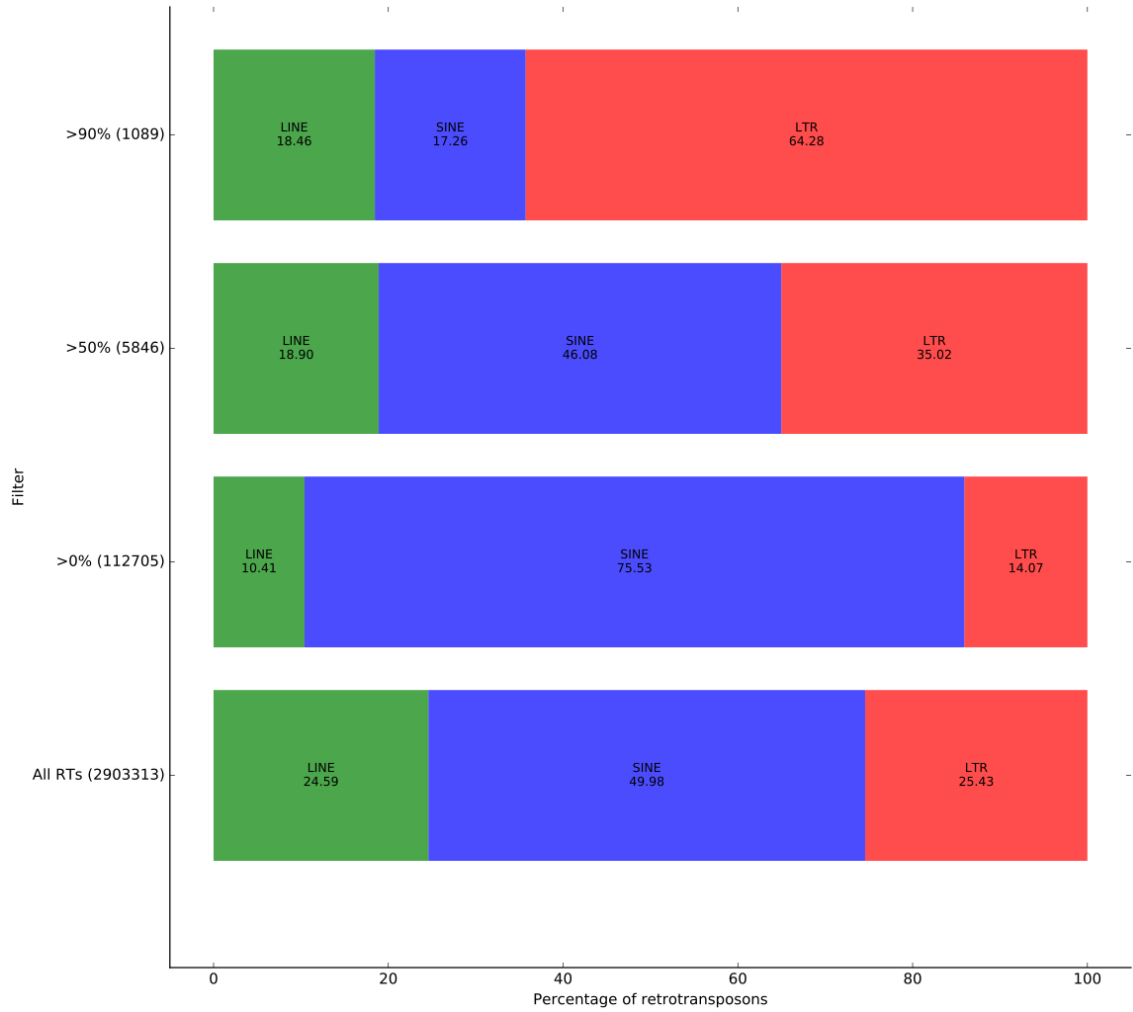


Figure 5.4: The proportion of each of the major retrotransposon classes with different filters applied. All RTs: all retrotransposon (RT) elements in the genome; >0%: all RTs overlapping a reconstructed transcript; >50%: RTs overlapping a reconstructed transcript with more than 50% RT content; >90%: RTs overlapping a reconstructed transcript with more than 90% RT content. As more stringent filters are applied, LTRs become enriched compared to their proportion across the genome, while SINEs are depleted.

5.2 Effect of RT transcripts on gene expression

As discussed in the Introduction, there is a well-established link between lncRNA and retrotransposons, with many lncRNAs containing retrotransposon sequence.

Their relationship with protein-coding genes is less clear, if any such relationship exists. There are multiple examples of retrotransposon regulatory sequence being incorporated into gene regulatory networks (e.g., the interferon pathway [99]), and some genes have evolved from retrotransposons (e.g., syncytin [132]; see Introduction). These examples aside, the majority of published studies in this field focus on the relationship between retrotransposons and lncRNAs.

The presence of retrotransposon sequence in protein-coding transcripts would allow for regulation based on sequence similarity with the retrotransposon. This could be at the transcriptional level (e.g., targeting of epigenetic marks), or at the post-transcriptional level, through RNA-based mechanisms [246–248]. In this section I will quantify how the RT-containing transcripts already identified correspond to annotated protein-coding transcripts using the Ensembl annotation (see Methods). I will then test whether RT-containing transcripts affect the expression of protein-coding genes in *cis*.

To quantify the relationship between protein-coding transcripts and the RT-containing transcripts identified in the previous sections, I used the gffcompare tool (see Methods). I applied this tool to the RT-containing transcripts, so they would be classified based on their correspondence to a protein-coding transcript. The results are shown in Figure 5.5. As the proportion of retrotransposon sequence in the query transcripts increases, the proportion corresponding to a protein-coding gene decreases.

In order to assess whether retrotransposons in ncRNAs were affecting gene expression, I used data from three kinds of novel transcripts identified by the gffcompare program: intronic, antisense, and intergenic. For intronic and antisense transcripts, there is a corresponding reference transcript. For intergenic tran-

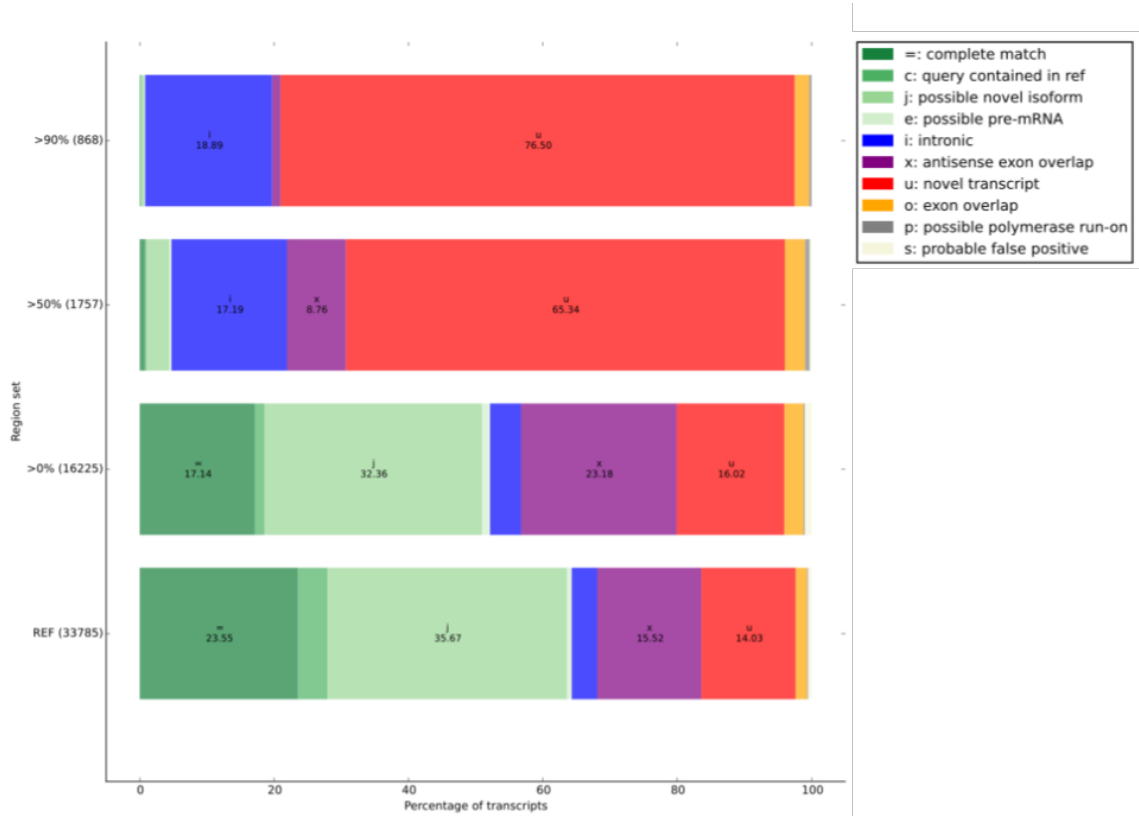


Figure 5.5: The proportion of reconstructed transcripts in each gffcompare category, for different levels of retrotransposon sequence content. Transcripts matching protein-coding reference transcripts tend to have lower retrotransposon content, as expected, and novel transcripts, which are potential lncRNAs, tend to have higher retrotransposon content.

scripts, one can find a corresponding reference transcript by looking in a window around the novel transcript. For each kind of novel transcript, I divided the novel transcripts into two categories: those with retrotransposon content, and those without. I then calculated the distribution of Spearman’s rho correlation values between the novel transcripts and their corresponding reference transcripts (see Methods), and used an Anderson-Darling (AD) test to compare the distributions statistically.

Figures 5.6, 5.7, 5.8 show the results of this analysis for intronic, intergenic,

and antisense novel transcripts. Table 5.3 shows the results of AD tests. In the first two cases, there is no clear bias towards high or low correlation values, and no difference between transcripts with or without retrotransposon content; this is confirmed by the AD test. The intronic transcripts with RTs show a slightly different distribution to the other intronic transcripts, with a small peak at around $\rho = -0.5$, but it is not statistically significant.

For antisense transcripts, there is a clear bias towards high correlation values, and antisense transcripts with retrotransposon content tend to have higher correlation values than those without; again, this is confirmed by the AD tests. The high correlations in expression between antisense transcripts and the protein-coding transcripts they overlap could be interpreted as a consequence of transcriptional activity in that region: if the gene is being expressed, then by chance an antisense transcript is also expressed. If this were the case, however, similar distributions would be expected for the intronic and intergenic transcripts as well. This is not the case, suggesting that the high correlation values may be meaningful.

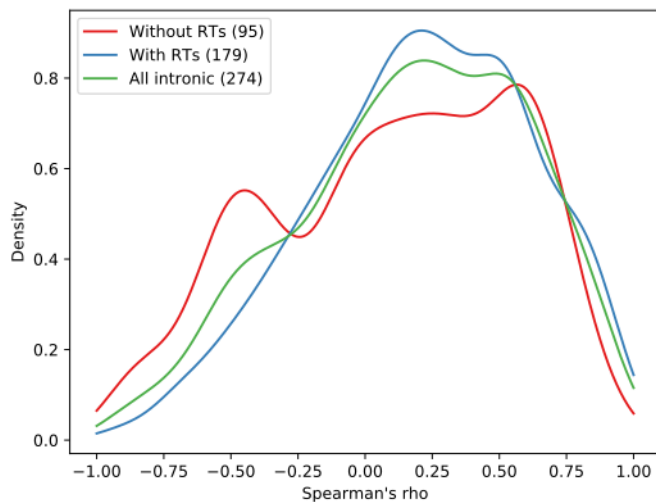


Figure 5.6: The distribution of correlation coefficients between expression levels of intronic StringTie transcripts and their corresponding protein-coding transcripts.

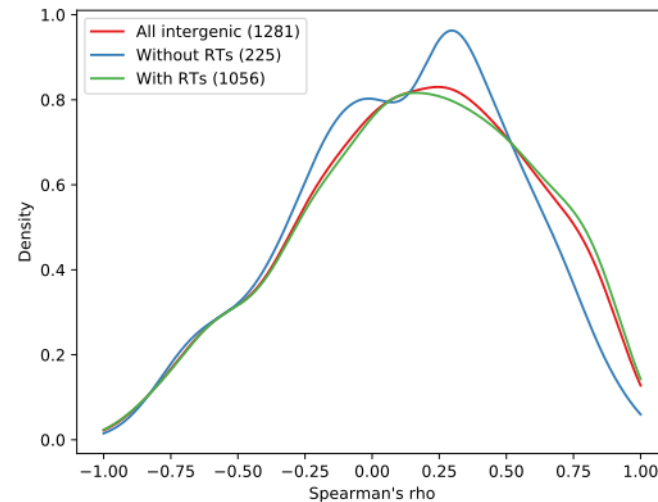
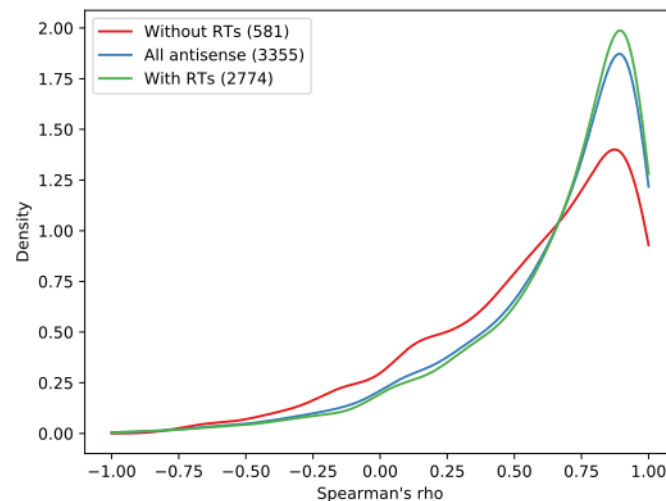


Figure 5.7: The distribution of correlation coefficients between expression levels of intergenic StringTie transcripts and protein-coding transcripts within 5kb.

Figure 5.8: The distribution of correlation coefficients between expression levels of antisense StringTie transcripts and their corresponding protein-coding transcripts.



	Antisense	Intronic	Intergenic
All vs. With RTs	-1.31	-0.65	-1.04
All vs. Without RTs	27.32	0.13	0.74
With RTs vs. Without RTs	27.32	1.76	1.54
Significance level	5%	2.5%	1%
Critical value	1.96	2.72	3.75

Table 5.3: Upper table: Anderson-Darling (AD) statistic values comparing the distributions of Spearman’s rho values shown in Figures 5.6, 5.7, and 5.8. Lower table: critical values for the AD statistic at different significance levels. For intronic and intergenic transcripts, none of the comparisons have AD statistics exceeding any of the critical values, and there is no evidence to suggest that any of their distributions are significantly different. For antisense transcripts, the distribution for the “Without RTs” category is significantly different from both other categories. These results are consistent with visual inspection of the distributions, and confirm that the presence of RTs correlates with higher Spearman’s rho values in sense/antisense pairs.

To further investigate whether antisense transcripts are indeed regulating the protein-coding transcripts they overlap with a retrotransposon-based mechanism, then two additional observations are of interest:

- The exact type of retrotransposon present in the antisense transcript
- Whether the protein-coding transcript also contains retrotransposon sequence

Figure 5.9 shows the retrotransposon content of antisense transcripts. The clusters are noisy, with a mixture of retrotransposon types, and relatively low retrotransposon content overall. This is not surprising, given that they overlap coding genes on the opposite strand, which are unlikely to contain large retrotransposon sequences. There is no apparent bias towards any single type of retrotransposon.

Figure 5.10 shows the distribution of correlation values between antisense

transcripts and their corresponding protein-coding transcripts, but with the RT-containing antisense transcripts divided into two groups based on the retrotransposon content of the protein-coding transcript:

- RTs match: the antisense transcript and the corresponding protein-coding transcripts both contain sequence from the same kind of retrotransposon
- RTs mismatch: the antisense transcript contains RT sequence, but the protein-coding transcript does not, or contains sequence from different kinds of RTs

In this case, there is a stronger bias towards higher correlation values if the antisense transcript and the protein-coding transcript both contain the same kind of retrotransposon content, compared to the other categories. This suggests that the presence of retrotransposon sequence may contribute towards regulation through sequence identity. If this is the case, the effect appears to be positive, as the correlations are exclusively positive. Antisense transcripts in general have the potential to regulate corresponding sense transcripts via sequence identity, which may account for the high correlations seen across all antisense transcripts. However, the antisense/sense pairs that share a retrotransposon type are more likely to have high correlation values, suggesting that the shared retrotransposon sequence may be better suited to such a mechanism. Indeed, it may be the case that both transcripts overlap the same retrotransposon element.

This supports previous findings linking antisense transcripts to gene regulation, and suggests that this is usually positive regulation. In addition, it suggests that shared retrotransposon sequence may facilitate such interactions.

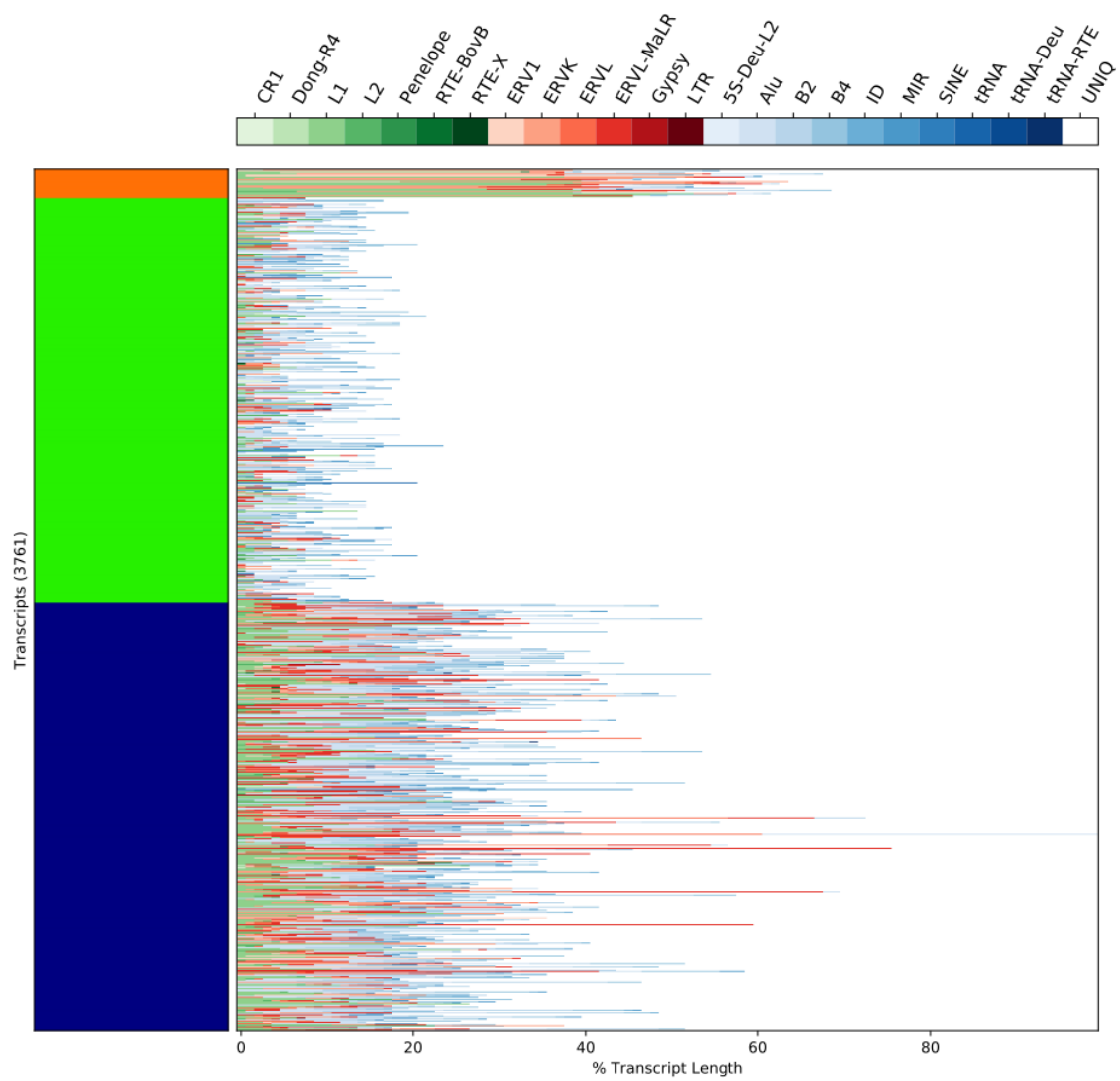


Figure 5.9: The retrotransposon content of antisense transcripts.

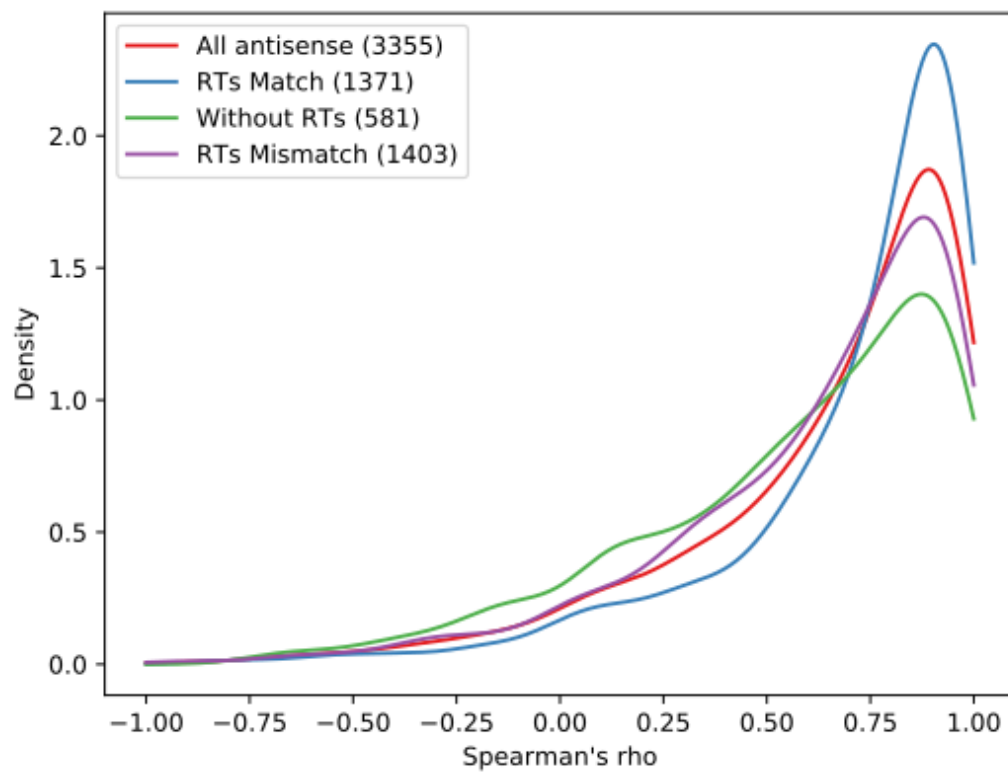


Figure 5.10: Expression correlation distributions for antisense transcripts in each retrotransposon matching category.

	Antisense
All vs. RTs Match	13.40
All vs. RTs Mismatch	12.90
All vs. Without RTs	27.32
RTs Match vs. RTs Mismatch	42.00
RTs Match vs. Without RTs	52.15
Without RTs vs. RTs Mismatch	6.08

Table 5.4: Anderson-Darling statistic values comparing the distributions of Spearman’s rho values shown in Figure 5.10. All of these values exceed the critical value for 1% significance, indicating that all four distributions are significantly different from each other. In particular, the distribution of values for the “RTs Match” category is very different from the distributions for the “RTs Mismatch” and “Without RTs” categories.

5.3 Cell-type Specificity

Previous studies [11,13,14] have observed cell-type specificity in lncRNA transcription, and I hypothesised that retrotransposon transcription in B and T cells would also show specificity. I also hypothesised that, as part of the same cell lineage, B and T cells would have a more similar retrotransposon transcription profile compared to another cell type. For this comparison I used an RNA-seq dataset from liver, as described in Datasets.

Upon visual inspection, the retrotransposon transcription profiles of B, T, and liver cells appear similar, all containing the large noisy cluster with low retrotransposon content, and a small cluster of high retrotransposon content transcripts. The primary visible difference is an extra cluster in T cells of high-ERV1 transcripts; however, it may be that similar transcripts exist in B and liver cells, but have not been separated from other high retrotransposon clusters (relevant figures can be found in Online Resources).

After filtering out transcripts with less than 50% retrotransposon content, B, T, and liver all show new clusters, similar to Figures 5.2 (full sets of figures can be found in Online Resources). These clusters appear similar, and so to quantify this I performed a cluster comparison, as described in Methods (Figure 5.11). This analysis confirmed that there are similar clusters of retrotransposon-containing transcripts in both B and T cells; in particular, high LTR content transcripts and high LINE content transcripts. When compared to the liver transcripts, I found similar results when comparing B against liver and T against liver (see Online Resources).

From this, I concluded that at a broad level there are similar groups of retrotransposon-

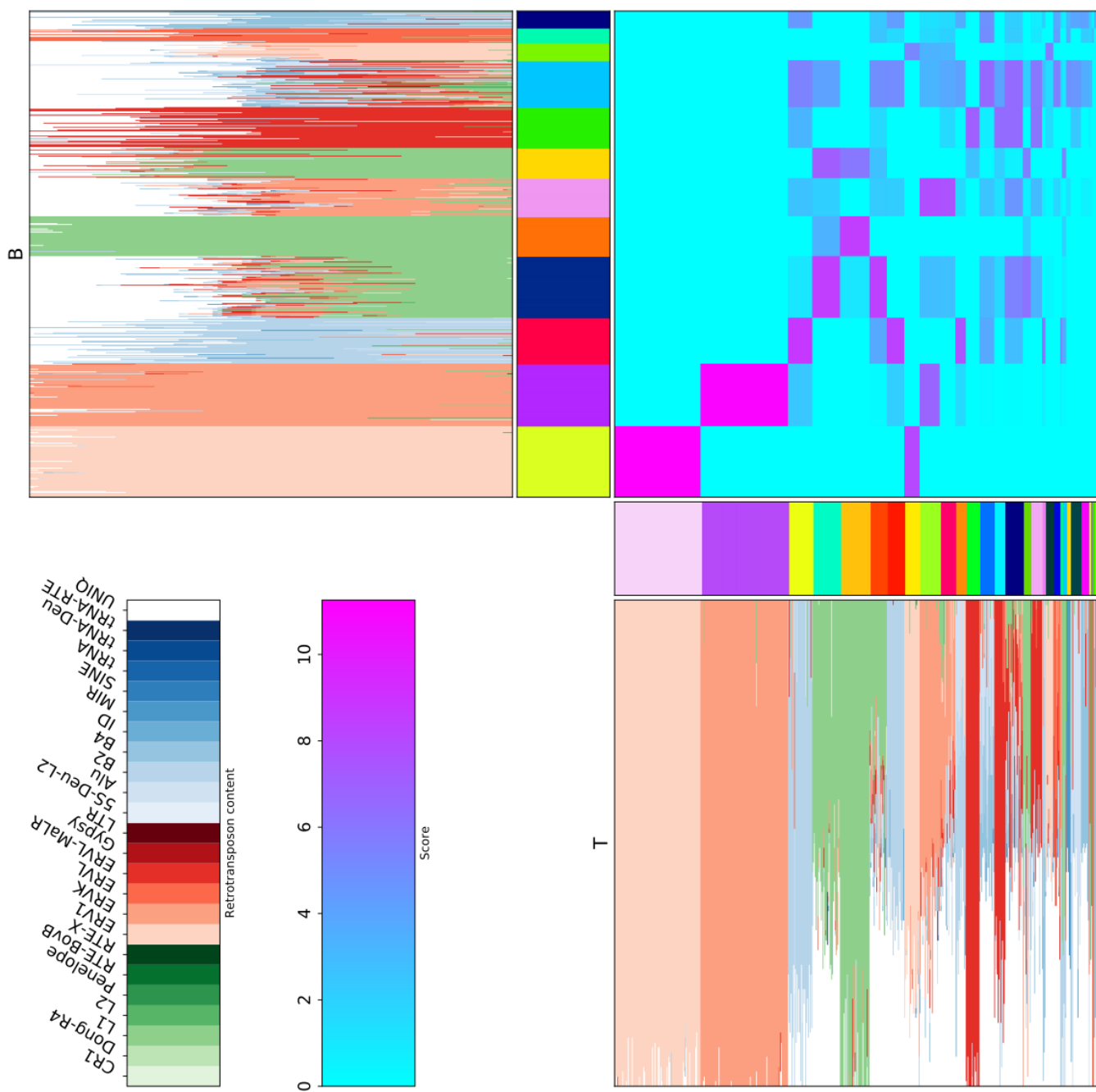


Figure 5.11: Cluster comparison between B and T cells based on retrotransposon content of transcripts with more than 50% retrotransposon sequence. The central heatmap shows the similarity score between each pair of B/T clusters; a higher score means the clusters are more similar. The clusters are also shown, in the same manner as described in Methods and shown in Figure 5.2. This confirms the existence of similar clusters of transcripts with high retrotransposon content in both B and T cells.

containing transcripts in B, T, and liver cells, including transcripts with high L1 and ERV content. However, it may be that the individual retrotransposons being transcribed in each case are specific to each cell type. Figure 5.12 shows the number of individual retrotransposon elements shared between the cell types, with different content level filters applied. As more stringent filters are applied, the individual retrotransposon elements tend to be cell-type specific (Tables ??). This would be consistent with the idea that transcripts with lower retrotransposon content tend to be coding transcripts with non-functional retrotransposon content, which are more likely to be shared between cell types. In contrast, those with higher retrotransposon content may be functional lncRNAs that make use of the retrotransposon content, displaying the characteristic tissue specificity.

Label	B	T	L	B, T	B, L	T, L	B, T, L
Observed	2,323	2,864	1,260	1,336	40	40	74
Expected	2,592	1,889	783	2,192	103	95	283

$$\chi^2 = 1,381.59$$

$$p = 2.35 \times 10^{-295}$$

Table 5.5: The results of a chi-squared test comparing the number of retrotransposons in each Venn category for all transcribed retrotransposons (expected), and for those in a transcript consisting of more >50% retrotransposon content (observed) (see Methods). The results are significant, and the observed versus expected values suggest that with the 50% filter the individual retrotransposons are more likely to be cell-type specific.

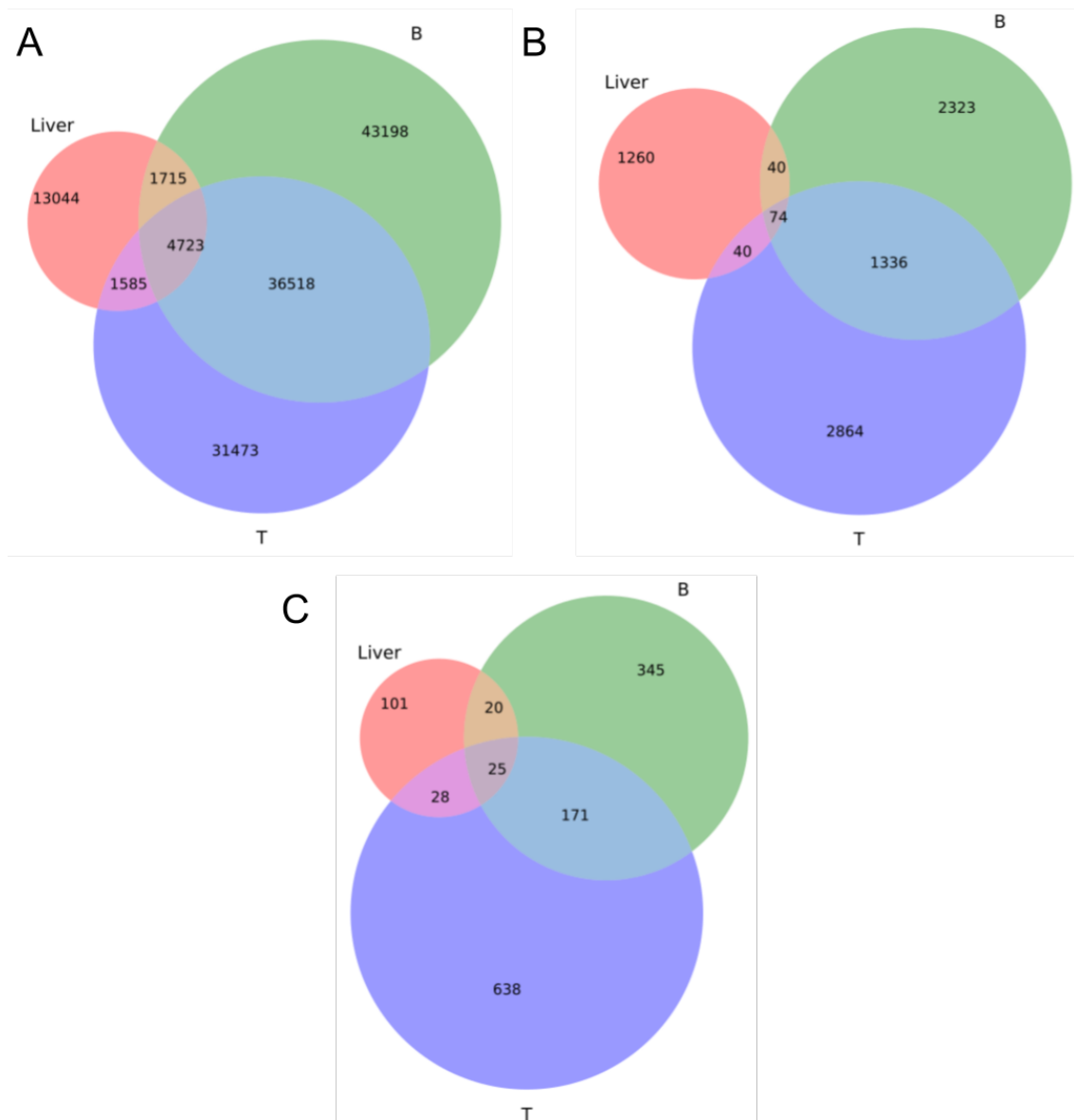


Figure 5.12: Overlap in transcribed retrotransposons between liver, B cells, and T cells. (A) All transcripts with retrotransposon content. (B) Transcripts with >50% retrotransposon content. (C) Transcripts with >90% retrotransposon content. In all three cases, there is a higher degree of shared retrotransposons between B and T than with liver, suggesting lineage-specificity as well as cell-type specificity.

Label	B	T	L	B, T	B, L	T, L	B, T, L
Observed	345	638	101	171	20	28	25
Expected	389	479	211	224	7	7	12

$$\chi^2 = 234.24$$

$$p = 9.51 \times 10^{-48}$$

Table 5.6: The results of a chi-squared test comparing the number of retrotransposons in each Venn category for retrotransposons in a transcript consisting of more >50% retrotransposon content (expected), and those in a transcript consisting of more >90% retrotransposon content (observed) (see Methods). The results are similar to those shown in Table 5.5, suggesting that retrotransposons are more likely to be cell-type specific with the more stringent filter.

5.4 Summary

In this chapter, I have quantified the retrotransposon content of B and T lymphocytes, and demonstrated that a significant proportion of transcripts contain some degree of retrotransposon sequence. Most of these transcripts contain only a small amount of retrotransposon sequence, but a small percentage contain a high proportion of RT sequence. These are enriched for ERVs and L1s, and tend to be intergenic transcripts that do not correspond to a known protein-coding transcript; however, their function cannot be ascertained from this analysis. Transcripts of this kind are found in B, T, and liver cells, but usually do not represent the same individual retrotransposon elements. It may be that there is a common role for high-retrotransposon sequences in multiple cell types, but different retrotransposons fulfil this role under different epigenetic conditions (e.g., chromatin conformation).

A minority of the reconstructed transcripts can be classified as antisense to a known protein-coding transcripts. In general, there are positive correlations between expression of the antisense transcript and the corresponding protein-coding transcript; however, in this chapter I have shown that these correlations are stronger if the antisense transcript contains retrotransposon sequence, and even more so if the protein-coding transcript contains matching retrotransposon sequence. Antisense transcripts have the potential to form RNA:RNA duplexes with the corresponding sense transcripts, if there is sufficient corresponding sequence between the two. It may be that the presence of retrotransposon sequence in both facilitates the formation of such duplexes, and that this protects the mRNA from degradation.

Chapter 6

Retrocopy Transcription in Mouse Lymphocytes

Retrocopy transcription is now known to be widespread, giving rise to non-coding transcripts that represent an indirect contribution of retrotransposons to the transcriptome in mammals [141,159,161]. Recent work has suggested that a retrocopy RNA may regulate the transcript from which it was copied [160,172], possibly through mechanisms based on sequence identity (see Introduction for a more detailed review). However, this is still an open question, and relatively few studies have explored this. The BLUEPRINT RNA-seq datasets provided an excellent opportunity to do so.

In this chapter, I hypothesise that:

- retrocopy lncRNAs (RC-lncRNAs) do affect the expression of the genes from which they originate
- this is a regulated, functional effect

- this occurs through a mechanism based on the sequence similarity between the RC-lncRNA and the parent mRNA

To investigate these hypotheses, I will focus only on the BL6 datasets initially, as there are high quality annotations for both reference genes and retrocopies. I will also use the CAST datasets to investigate possible conservation of retrocopy transcripts, but due to the lack of high quality annotations for the CAST genome, I was not able to conduct an equivalent analysis in CAST. Any future work on this topic should include the creation of a high quality retrocopy annotation for CAST, if one does not exist, and a more extensive analysis of retrocopy expression in CAST.

6.1 Retrocopies are Expressed in Mouse Lymphocytes

In order to discover expressed retrocopies in the RNA-seq samples, I compared the reconstructed transcripts to annotated retrocopies using a similar approach to that used with retrotransposons (see Methods). I added an additional step to retain only transcripts with a high degree ($>80\%$) of reciprocal overlap with a retrocopy.

Using this method, I discovered 994 expressed retrocopies across all of the BLUEPRINT BL6 RNA-seq samples, corresponding to 1,073 transcripts. The 994 expressed retrocopies originated from 456 parent transcripts. 1,010 of the transcripts were novel transcripts, with no corresponding annotation in ENSEMBL, while 54 matched a previously annotated transcript to some degree (Table 6.1).

The expressed retrocopies and their parent transcripts are distributed fairly

Code	Description	Number
=	Complete match	9
c	Query contained in reference	1
e	Possible pre-mRNA	7
j	Possible novel isoform	20
o	Exon overlap	17
p	Possible polymerase run-on	9
i	Intronic	164
u	Novel transcript	734
x	Antisense exon overlap	112

Table 6.1: Comparison of retrocopy transcripts to Ensembl reference transcripts according to gffcompare, across all BL6 samples. Codes “i”, “u” and “x” can be regarded as novel transcripts (highlighted in blue, 1,010 in total), while the remaining codes indicate a type of match. However, only 9 match exactly (code “=”).

evenly throughout the genome, with some hotspots (Figure 6.1). The distribution of expressed retrocopies across chromosomes follows that of all retrocopies (i.e., including both expressed and non-expressed), in general (Figure 6.2). Two notable exceptions are chromosome 6 and chromosome Y. The distribution of expressed retrocopy parents follows the distribution of all retrocopy parents fairly well, with no notable outliers.

Chromosome 6 is noticeably enriched for expressed retrocopies. A gene ontology (GO) analysis of genes on chromosome 6 showed an enrichment for genes involved in immunoglobulin production and the immune response (Table 6.2). It may be that the large regions of chromosome 6 are in an open conformation so that immune response genes are expressed or ready for expression in lymphocytes. As a result of this, retrocopies are more likely to be transcribed in this more permissive

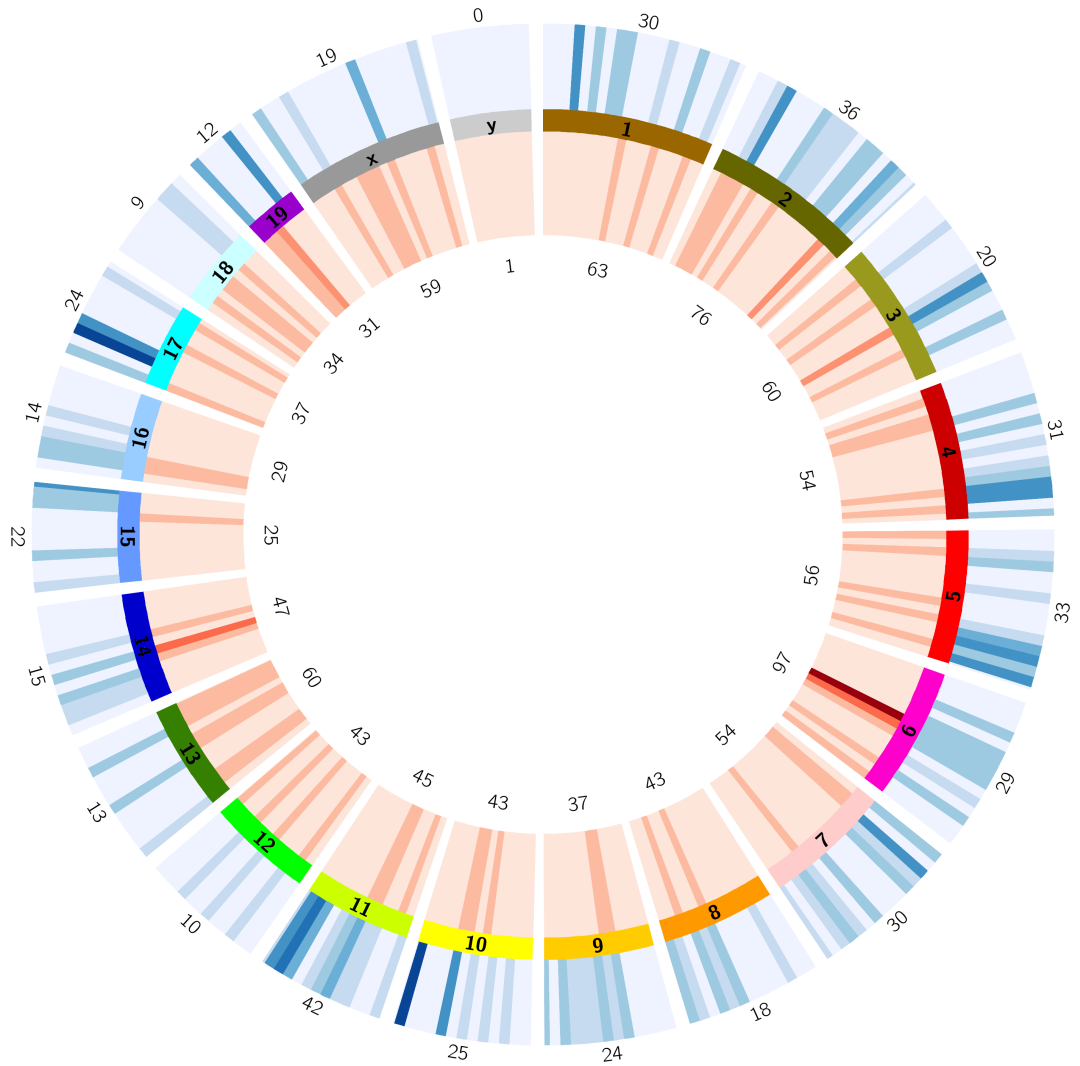


Figure 6.1: Distribution of expressed retrocopies (inner red heatmap, inner numbers) and their parent transcripts (outer blue heatmap, outer numbers) in the mouse genome. Created using the Circos software [238].

environment.

Chromosome Y is almost completely depleted for expressed retrocopies, whereas across all retrocopies, the greatest proportion are on chromosome Y. This is probably the result of two features of the Y chromosome and its evolution. Firstly, the

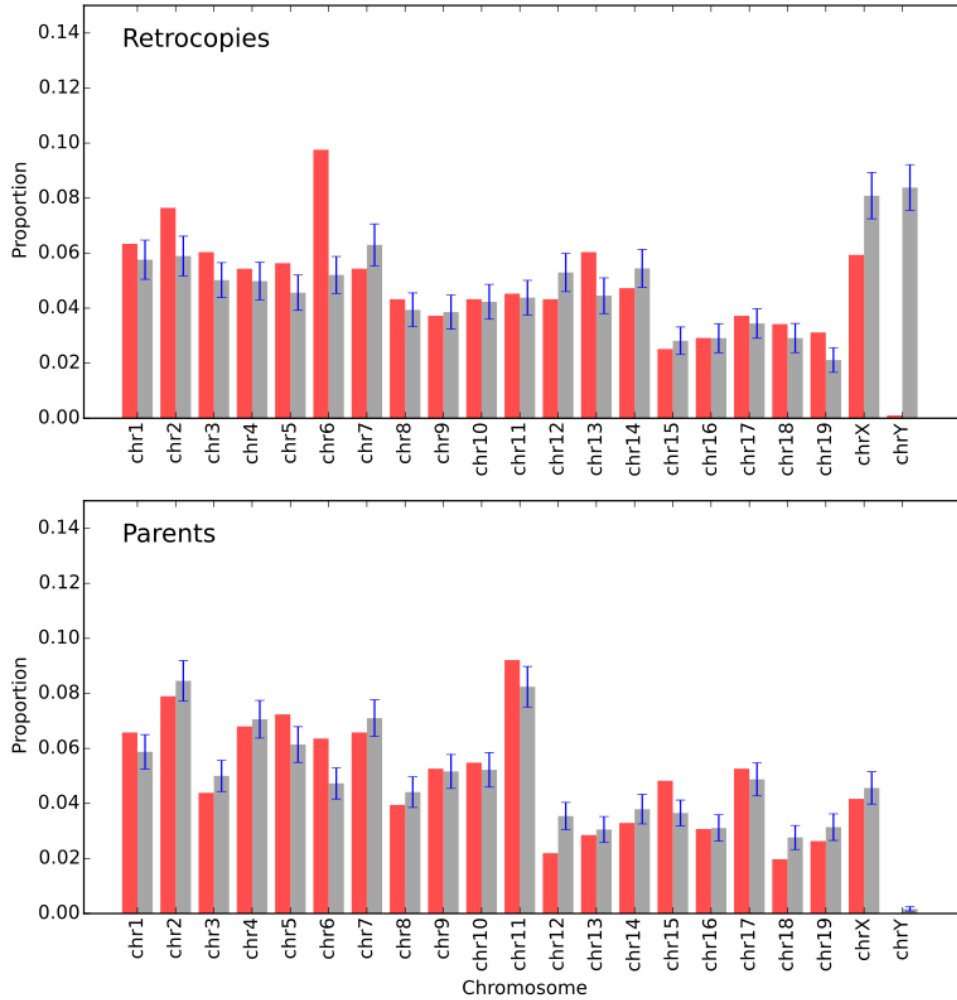


Figure 6.2: The proportion of retrocopies and their parents found on each chromosome, both expressed (red) and randomly selected (grey). Error bars indicate the standard deviation of proportions from 1000 random samples, each of similar size to the number in the expressed set.

mammalian Y chromosome is almost exclusively involved in testis determination and spermatogenesis, with little or no function outside of these roles [250]. In particular, it does not play a known role in the immune response, and is there-

GO biological process	Ref	Query	Expected	Enrichment	+/-	p-value
detection of chemical stimulus involved in sensory perception of bitter taste	48	30	2.38	12.58	+	4.79×10^{-19}
immunoglobulin production	163	71	8.10	8.77	+	1.74×10^{-38}
→ production of molecular mediator of immune response	181	71	8.99	7.90	+	1.29×10^{-35}
→ immune system process	1,915	153	95.14	1.61	+	4.66×10^{-5}
response to pheromone	104	37	5.17	7.16	+	6.34×10^{-16}
immune response	1,087	107	54.01	1.98	+	2.91×10^{-7}

Table 6.2: The results of a gene ontology (GO) analysis of the genes on chromosome 6. There is significant enrichment for genes related to the immune response, suggesting that chromosome 6 will be in an open conformation in lymphocytes. GO analysis carried out using the online GO Enrichment Analysis tool from the Gene Ontology Consortium [249].

fore likely to be silenced in lymphocytes, leading to the observed depletion for retrocopy transcription. Secondly, its role in spermatogenesis suggests it will be in a more permissive chromatin conformation in male germ cells, making it a more likely target for retrocopy insertion. Such germline insertions are then fixed in the genome. In combination, this means that the Y chromosome has a relatively high number of retrocopies, but they are unlikely to be expressed in lymphocytes.

A GO analysis of the parent genes corresponding to the expressed retrocopies showed a number of enriched biological processes (see Online Resources). However, a GO analysis of all retrocopy parents showed enrichment for all of the same biological functions (see Online Resources), suggesting that no specific functional subset has been chosen for expression, and that retrogenes are simply more likely

to originate from certain groups of genes (e.g., germline-expressed genes, house-keeping genes).

6.2 Retrocopy expression is shared across lineages

Comparing transcript expression across cell types can be used to identify putative functional transcripts. Expression of a given transcript specifically in one cell type may indicate a specific role for that transcript in that cell type; transcripts shared across multiple diverged cell types may be involved in a pathway common to many cell types.

To investigate these possibilities for retrocopy transcripts, I ran the retrocopy discovery pipeline on samples from each cell type separately, to obtain sets of expressed retrocopies for B cells and T cells. By comparing these two cell types, I obtained lists of cell type-specific and shared expressed retrocopies, and retrocopy parents. The majority of expressed retrocopies and their parents were found in both B and T cells, with a small number showing cell-type specificity (Figure 6.3). A statistical analysis (see Methods and Online Resources) showed a significant enrichment towards shared retrocopies ($p < 4.94 \times 10^{-324}$) and parents ($p < 4.94 \times 10^{-324}$), with a corresponding depletion in cell type-specific retrocopies and parents. The sets of parents with expressed retrocopies found in one cell type or the other were not enriched for any biological functions, according to a gene ontology analysis. Those shared between both cell types showed the same enrichments as shown by all retrocopy parents, expressed or otherwise.

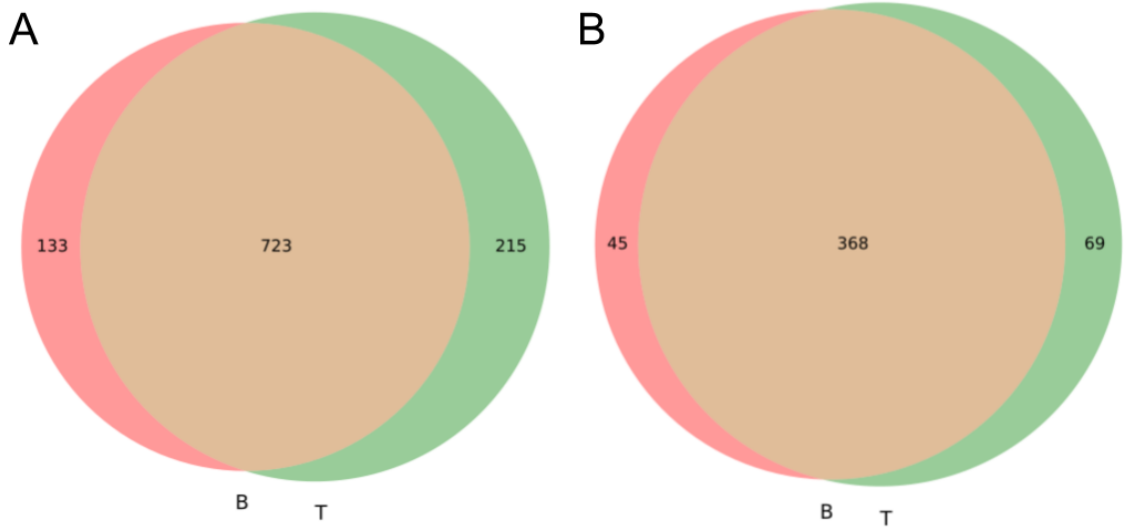


Figure 6.3: (A) Retrocopies expressed in B and T cells. (B) Parents of retrocopies expressed in B and T cells. In both cases, there is a high degree of overlap between the cell types.

B and T cells are relatively close cell types, both being derived from lymphoid progenitor cells. The high proportion of shared expressed retrocopies and parents could be a reflection of their shared lineage. To investigate this, I used publicly available RNA-seq data from mouse liver as an outgroup (see Datasets). I applied the same analysis pipeline to obtain a set of expressed retrocopies in liver (Table 6.3). Fewer retrocopies are expressed in the liver, which may be a reflection of the small number of samples available in that dataset, although they have higher sequencing depth than the BLUEPRINT samples. A better comparison could be obtained by downsampling the BLUEPRINT data before comparing to the liver samples; such a comparison should be included in future work.

I compared the liver results to the results from the B and T cells (Figure 6.4). For both retrocopies and their parents, there is significant enrichment in the number shared between all three cell types ($p < 4.94 \times 10^{-324}$ for both retrocopies

Transcript Set	Retrocopy Transcripts	Retrocopies	Parent Transcripts
ALL	1,073	994	456
B	901	856	413
T	979	938	437
Liver	375	348	253

Table 6.3: Number of retrocopy transcripts, expressed retrocopies, and corresponding parent transcripts across all BLUEPRINT samples, in B cells, in T cells, and in liver.

and parents) and for the number shared between B and T cells but not liver ($p < 4.94 \times 10^{-324}$ for both retrocopies and parents). There is a depletion in retrocopies and parents found in one cell type only. This suggests that while retrocopy expression does reflect cell lineage, there is also a core set of retrocopies and parents that are expressed across diverged cell types. This result is stronger in the parents than in retrocopies.

This analysis does not directly suggest a functional role for these retrocopies. The different sets of retrocopies that are expressed could be an effect of chromatin conformation, so retrocopies are expressed by chance when other transcripts nearby are expressed. The analysis of parent genes is more interesting, as a single parent transcript can have multiple retrocopies in different areas of the genome, and so their inclusion in a given set is not necessarily linked to chromatin conformation and location. The evidence for a set of parents with expressed retrocopies shared across cell types is stronger than for individual retrocopies. It may be that certain transcripts are regulated using a retrocopy-based mechanism, and different specific retrocopies are used to achieve this in different cell types, depending on chromatin conformation or other regulatory factors. Future work on this should expand this

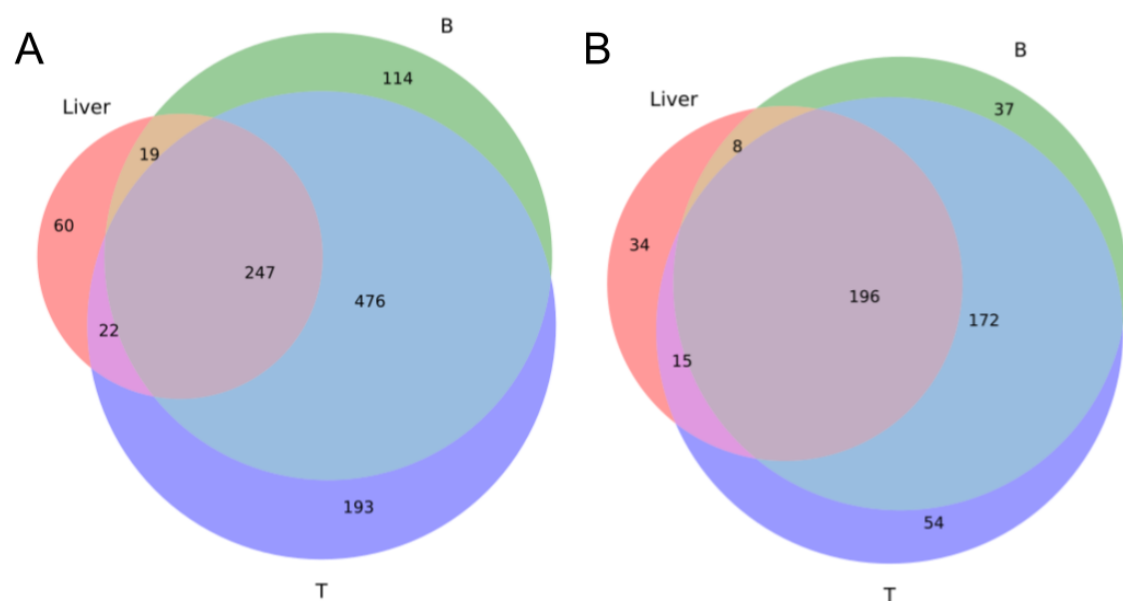


Figure 6.4: (A) Expressed retrocopies in B cells, T cells, and liver. (B) Parents of expressed retrocopies in B cells, T cells, and liver. Liver shows little overlap with B and T individually, but there is significant overlap between all three.

analysis to a range of other cell types; if this hypothesis is correct, we would expect to see a significant number of shared parents across multiple cell types.

6.3 Expressed Retrocopies Produce RNA Complementary to Their Parent

As noted above and in the Introduction, retrocopy transcripts have regulatory potential based on sequence similarity with their parent gene. The specific mechanism depends on whether the retrocopy RNA is complementary to the parent or not. If it is complementary, the retrocopy and parent RNAs can form RNA:RNA duplexes [172]; if not, the retrocopy RNA can act as an miRNA sponge, for example [168, 169]. Establishing a preference for one option or the other would rule out

one set of mechanisms in favour of the other.

The relative strandedness of the retrocopy RNA can be used to establish whether it is complementary to its parent transcript. When considering transcribed retrocopies and their relationship with their parent transcripts, there are three levels of strandedness (Figure 6.5):

- The strand of the parent transcript, i.e., the strand from which the parent transcript is transcribed
- The strand of the retrocopy annotation, which reflects the orientation of the insertion
- The strand from which the retrocopy is transcribed

There are therefore eight possible combinations of strands, which can be divided into two groups: those that produce retrocopy RNA (rcRNA) complementary to their parent mRNA, and those that do not (Figure 6.5).

Each retrocopy transcript is associated with a retrocopy and a parent, and so I assigned each of the expressed retrocopies to one of the eight strand combinations. This showed that the vast majority of expressed retrocopies are transcribed in such a way as to produce rcRNA complementary to their parent mRNA (Table 6.4). To assess the statistical significance of this result I performed a chi-squared test comparing the observed number in each category to the expected number in each category. To obtain the expected number I counted the number of retrocopies falling into each of the four parent/retrocopy strand combinations. I used the proportions of observed retrocopy transcript strand to calculate the expected number in each of the eight parent/retrocopy/transcript categories. This calculation

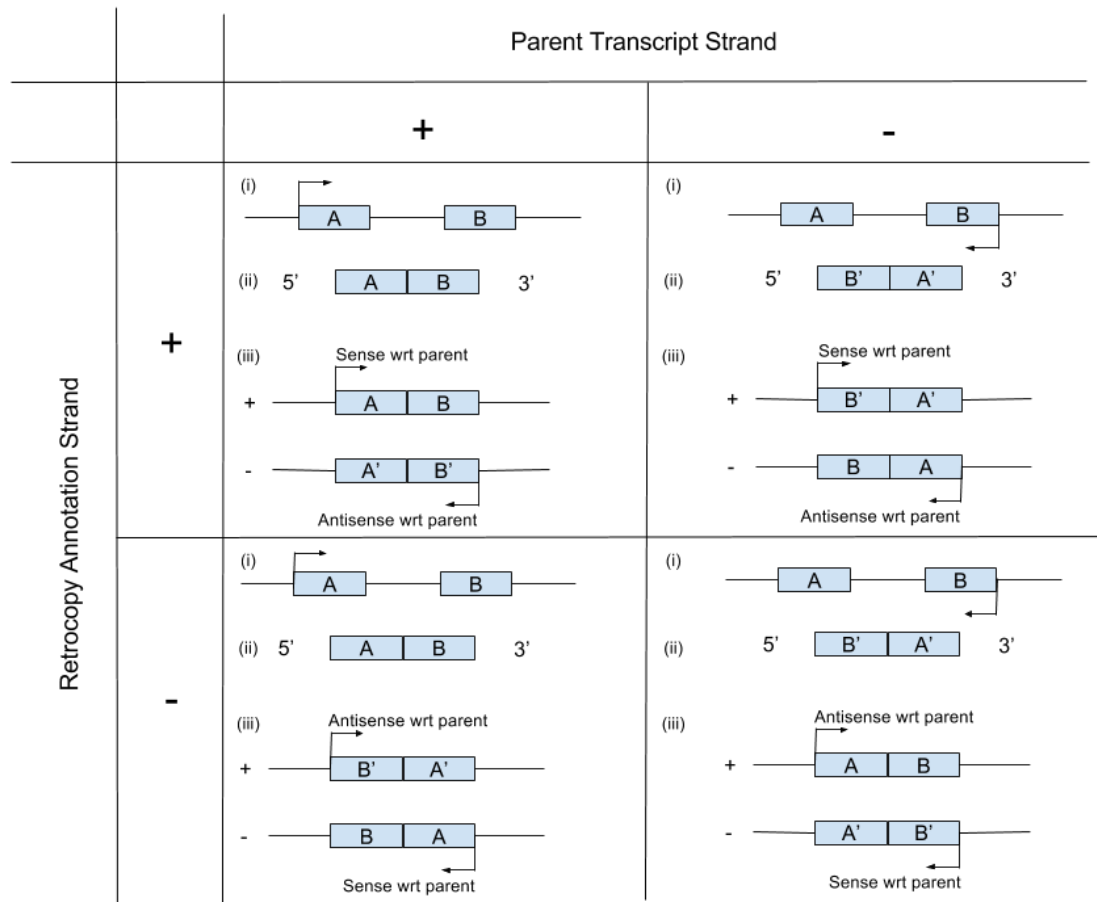


Figure 6.5: The possible combinations of parent transcript strand, retrocopy strand, and transcript strand. Blue lettered boxes represent exons. The addition of ' represents the reverse complement. (i) The original transcript in the genome. (ii) The parent mRNA. (iii) The retrocopy insertion and its possible transcription. Sense with respect to (wrt) the parent produces RNA equivalent to the parent. Antisense wrt the parent produces RNA complementary to the parent.

showed an essentially even distribution across all eight categories. The chi-squared test indicated that there is a significant bias towards combinations that produce antisense RNA. There is no bias towards any particular strand combination.

This suggests that any functional role played by the expressed retrocopies will tend to be based on RNA complementarity, rather than exact sequence identity.

Parent	Retrocopy	Transcript	Sense wrt parent	Count	Expected
+	+	+	Sense	11	138.08
+	+	-	Antisense	227	121.44
+	-	+	Antisense	288	138.02
+	-	-	Sense	12	121.38
-	+	+	Sense	8	137.49
-	+	-	Antisense	217	120.91
-	-	+	Antisense	256	149.33
-	-	-	Sense	39	131.33

Observed total antisense = 988

Observed total sense = 70

Expected total antisense = 529.70

Expected total sense = 528.28

$$\chi^2 = 809.67$$

$$p = 1.52 \times 10^{-170}$$

Table 6.4: The number of expressed retrocopies across all BLUEPRINT BL6 samples falling into each strand combination. A chi-squared test shows that there is a very clear enrichment in the categories leading to rcRNA complementary to the parent (highlighted in blue).

However, any such role would rely not just on complementarity from a strand perspective, but also on high sequence identity between the parent transcript DNA and the retrocopy DNA.

6.4 Expressed Retrocopies Have Higher Sequence Identity with Their Parents

The previous section demonstrates that the expressed retrocopies are much more likely to form RNA complementary to that of their parent transcript. If retrocopy expression is regulating parent transcripts, this bias rules out certain regulatory

mechanisms (e.g., miRNA sponge) in favour of others (e.g., formation of RNA:RNA duplexes). However, all of these mechanisms rely on a high level of sequence identity between the retrocopy and the parent. If the retrocopy has degraded over time, the sequence identity may be low, in which case it would not be able to fulfill a regulatory role based on sequence identity.

To assess the sequence identity between parents and retrocopies, I compared each retrocopy with its parent transcript and performed a local alignment between the two (see Methods). I assigned each alignment a score based on its length and identity, and plotted the distribution of scores for three sets of retrocopy/parent pairs: expressed retrocopies, non-expressed retrocopies, and a subset of retrocopies with a randomly assigned parent.

Figure 6.6 shows the results of this analysis. The distribution of alignment scores across all retrocopies shows a multimodal distribution, which could reflect the distribution of ages across the retrocopies. In this case, the leftmost peak would be the oldest retrocopies, which have decayed to the point that the alignment with the parent is no better than the alignment between the retrocopy and a random reference gene. In this case, the reliability of assigning a parent to a retrocopy seems dubious, if the alignment is no better than random; however, the score used here is a summary, and information is lost that could be used to identify a parent. This leftmost peak could represent a burst of retrocopy activity at a particular time, or a plateau of decay reached by most retrocopies eventually. The peaks to the right of the histogram could also be more recent bursts of retrocopy formation, as reported in [154, 251].

Overall, the expressed retrocopies tend to have higher levels of sequence identity with their parents compared to non-expressed retrocopies. It is therefore possible

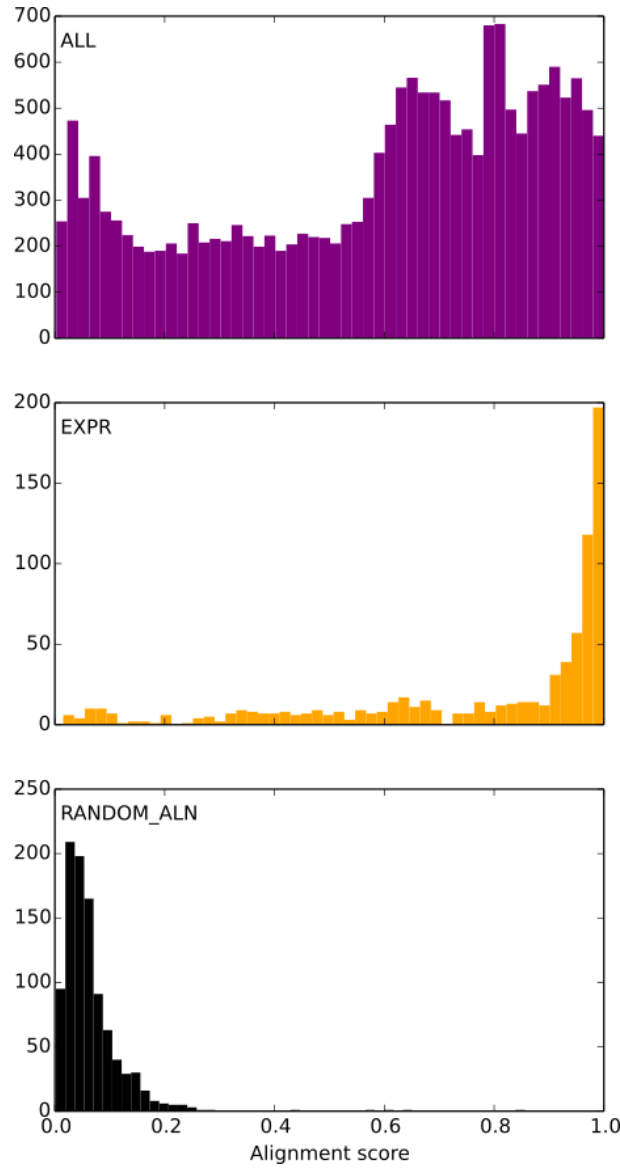


Figure 6.6: Retrocopy/parent alignment scores, where 1.0 represents a perfect full-length alignment (see Methods). ALL: All retrocopies. EXPR: All expressed retrocopies. RANDOM_ALN: Negative control where retrocopies are aligned to randomly chosen parent transcripts. Expressed retrocopies are clearly biased towards high scores compared to all retrocopies.

that the majority of the retrocopy transcripts identified here could interact with their parent transcripts based on shared sequence identity. Given the previously

described bias towards complementary rcRNA, I expected to observe a higher sequence identity among retrocopies producing complementary RNA. However, the number of retrocopies producing non-complementary RNA is too small to be reasonably compared with any other category (see Online Resources).

While this finding does not guarantee that expressed rcRNA regulates parent transcripts via a sequence-based mechanism, it allows for the possibility that such a mechanism exists. It should be noted that more recent retrocopies are expected to have a higher sequence identity with their parent, as they will have suffered fewer mutations. Indeed, methods similar to the alignment score used here are used to estimate the age of retrocopies [141]. Assuming that these retrocopy transcripts are functional, it may be that expressed retrocopies have undergone selection to preserve sequence identity with their parent, which is important for said function. Alternatively, it may be that expressed retrocopies are also younger retrocopies, and remain expressed until the sequence similarity to the parent has decreased to the point when it is no longer effective as a regulatory RNA. If this is the case, then how is their expression regulated? As described by Carelli *et al.*, expressed retrocopies either use pre-existing promoters, or evolve one *de novo*; the former case accounts for a small percentage of retrocopies, and the latter case requires time for such a promoter to evolve.

6.5 Retrocopy Transcription and Retrotransposon Indels

Over time, retrotransposons acquire mutations, including single nucleotide variations (SNVs) and structural variations (SVs). These can be deleterious, rendering the retrotransposon unable to successfully copy and paste itself. Amongst the SVs are “indels”: insertion of sequence not belonging to the retrotransposon, sometimes also accompanied by a deletion of retrotransposon sequence at the insertion locus (Figure 6.7). These have been identified by the RepeatMasker software by looking for fragments of the same retrotransposon near to each other, often with overlapping sequence on either side of the gap. 340,398 (11.7%) retrotransposons in the mouse genome contain at least one indel.

A naïve intersection of the reconstructed transcriptomes with retrotransposons (i.e., ignoring internal structures, see Methods) showed sets of transcripts with a high degree of overlap with retrotransposons, particularly L1 and certain ERV elements. Closer examination showed that some of these were not true retrotransposon transcripts, but instead transcripts from inside retrotransposon indels (RTIs) (Figure 6.7).

To quantify this, I compared the reconstructed transcriptomes with the full set of RTIs to find all transcribed RTIs. I also compared RTIs to the RepeatMasker annotation and the retrocopy annotation in order to classify each RTI based on its contents. Visual inspection of Figure 6.8 suggests that retrocopy-containing RTIs are highly enriched in the set of expressed RTIs compared to all RTIs, while other classifications are depleted. A chi-squared test showed that this is indeed the case (Table 6.5).

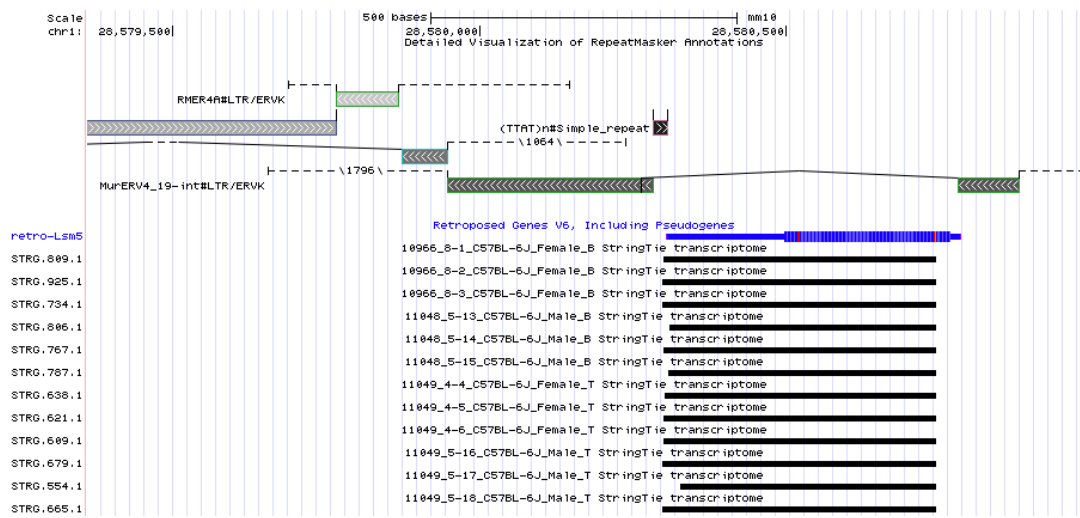


Figure 6.7: An example of a retrotransposon indel (RTI) with a retrocopy transcribed from inside it. A retrocopy of the Lsm5 gene has inserted into an ERVK element, creating an RTI with the retrocopy inside. The reconstructed transcriptomes show that this retrocopy is transcribed across all of the BL6 samples.

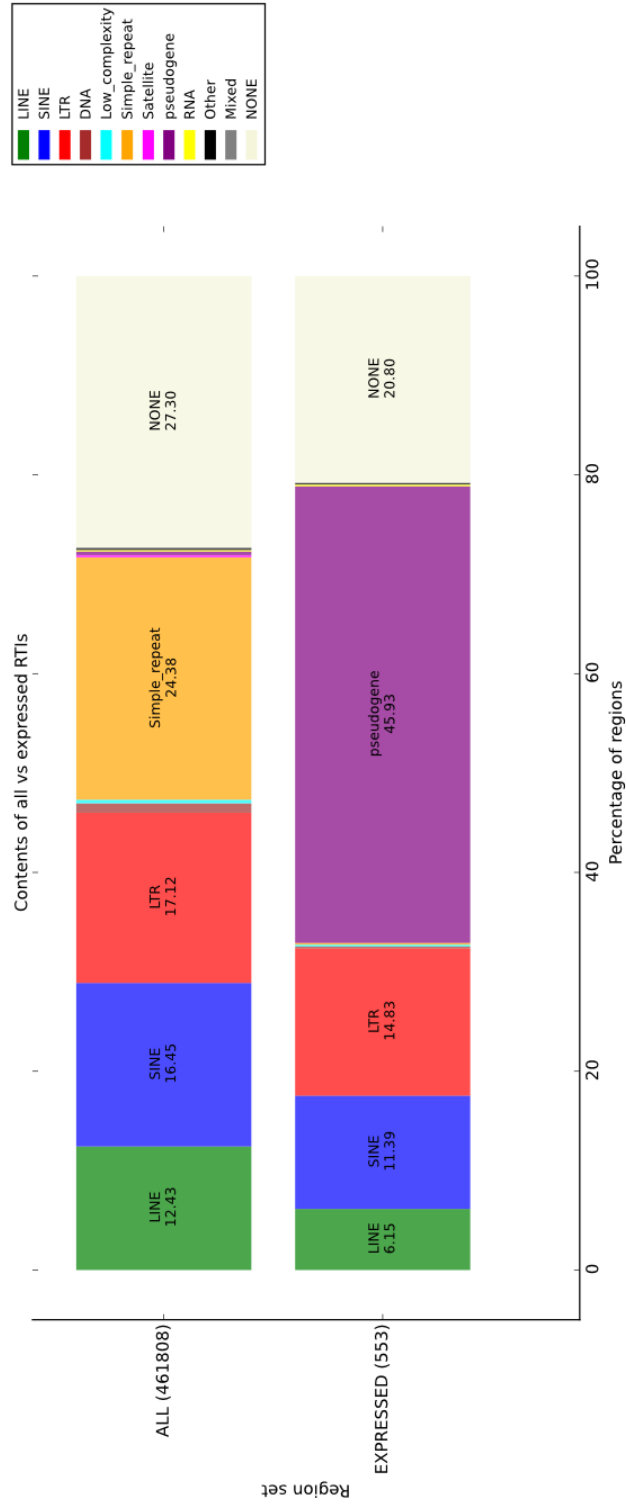


Figure 6.8: Classifications of RTIs by contents. In this case, “pseudogene” is synonymous with retrocopy. “NONE” means that the RTI could not be classified using repeats and retrocopies, and may contain other sequence. There is a significant enrichment for retrocopies in the expressed RTIs (Table 6.5).

Class	LINE	SINE	LTR	DNA	Low complexity	Simple repeat	Satellite
Observed	34	63	82	1	1	1	0
Expected	69	91	95	5	2	135	1

Class	Retro- copy	RNA	Other	Mixed	NONE
Observed	254	1	1	0	115
Expected	2	1	1	0	151

$$\chi^2 = 32,124.85$$

$$p < 4.94 \times 10^{-324}$$

Table 6.5: The results of a chi-squared test comparing the contents of all RTIs to those that are expressed. To obtain the expected values, the proportions falling into each category across all RTIs were multiplied by the total number of expressed RTIs. This shows a clear and significant enrichment for retrocopies in the expressed RTIs.

I then compared the set of expressed retrocopies to the full set of RTIs, and found 241 expressed retrocopies contained inside RTIs. While this is a minority of the expressed retrocopies, a chi-squared contingency test showed that there is a statistically significant enrichment for expressed retrocopies inside RTIs, with approximately double the expected number (Table 6.6).

	Inside RTI	Not inside RTI	
Expressed	241	753	
Not Expressed	2,002	15,460	
Total BL6 retrocopies			18,456

$$\chi^2 = 142.69$$

$$p = 6.857 \times 10^{-33}$$

Expected Values:

	Inside RTI	Not inside RTI	
Expressed	120.80	873.20	
Not Expressed	2,122.20	15,339.80	

Table 6.6: The results of a chi-squared contingency test comparing expression of retrocopies in BL6 and their location inside an RTI. The observed values differ significantly from the expected values, suggesting that retrocopy expression and location inside an RTI are not independent.

These results suggest a possible link between retrocopy expression and its position inside a retrotransposon. As noted previously, retrotransposons are a rich source of regulatory elements, and it may be that retrocopies have adopted retrotransposon promoters. It is not clear why having retrotransposon sequence at both ends should be important; it may not be, and future work should include a more comprehensive analysis of retrotransposons as regulatory elements for retrocopies.

6.6 Retrocopy Expression May Affect Parent mRNA Levels

If retrocopy transcripts are involved in the regulation of their parent transcripts, we would expect to see this reflected in the expression levels of the parent mRNA. To investigate this, I compared the expression of parent transcripts in B and T cells. Since a single parent transcript can give rise to multiple retrocopies, and each retrocopy has the potential to interact with the parent individually, I used cell-specific parents, rather than cell-specific retrocopies. I obtained a list of parents corresponding to expressed retrocopies for each cell type, and compared the two lists to produce three sets of parents: those with one or more retrocopies expressed across both cell types, those with at least one retrocopy expressed in B cells only (“B-specific parents”), and those with at least one retrocopy expressed in T cells only (“T-specific parents”). It should be noted that the first list includes parents with multiple distinct retrocopies expressed in each cell type separately. There are 45 B-specific parents, 69 T-specific parents, and 368 shared (Figure 6.3).

I used the Ballgown software to obtain fold change values for each Ensembl transcript between B and T cells, along with corresponding confidence values (see Methods). There is a visible bias towards upregulation of the parent when a retrocopy is expressed (Figure 6.9), although a only minority of parent transcripts show this pattern: 15 in B cells, and 19 in T cells (listed in Table 6.7).

As described above, expressed retrocopies tend to have higher sequence identity with their parents, and so it may be possible that retrocopy RNA to regulates parent transcript expression through a mechanism based on sequence identity. If this is the case, we might expect that the parent transcripts showing upregulation

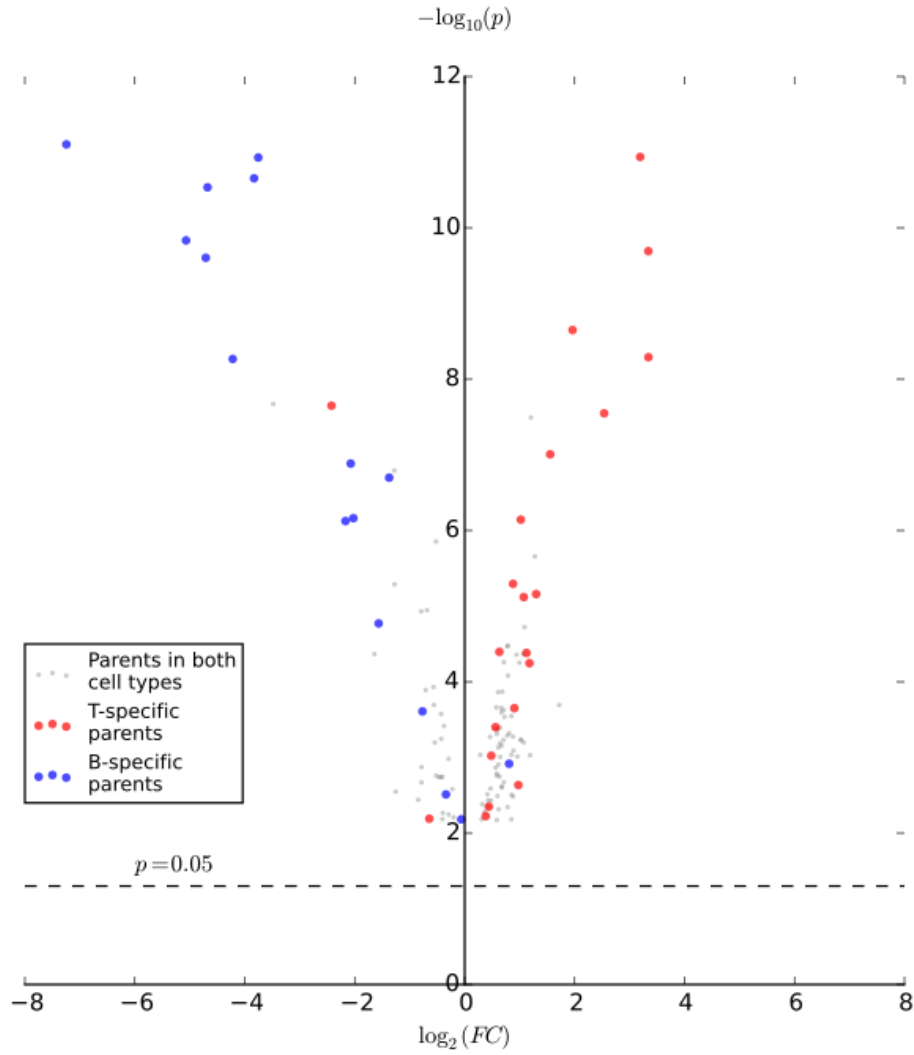


Figure 6.9: Fold change (FC) of retrocopy parents with a retrocopy expressed in either B cells or T cells. Positive log FC values indicate upregulation in T cells compared to B cells. There is a bias towards upregulation in the presence of a retrocopy in each case.

in the presence of an expressed retrocopy to have a higher level of sequence identity with these retrocopies. To investigate this, I obtained alignment scores for each expressed retrocopy corresponding to a cell type-specific parent with upregulation

in that cell type, and plotted these against the log-transformed fold change for each parent (Figure 6.10). This did not show any strong correlation between fold change and alignment score. While several of the parents used do have expressed retrocopies with high sequence identity, this is not true for all of the parents, and some have very low alignment scores. Even if only the retrocopy with the highest alignment score is used, there is still no correlation (see Online Resources).

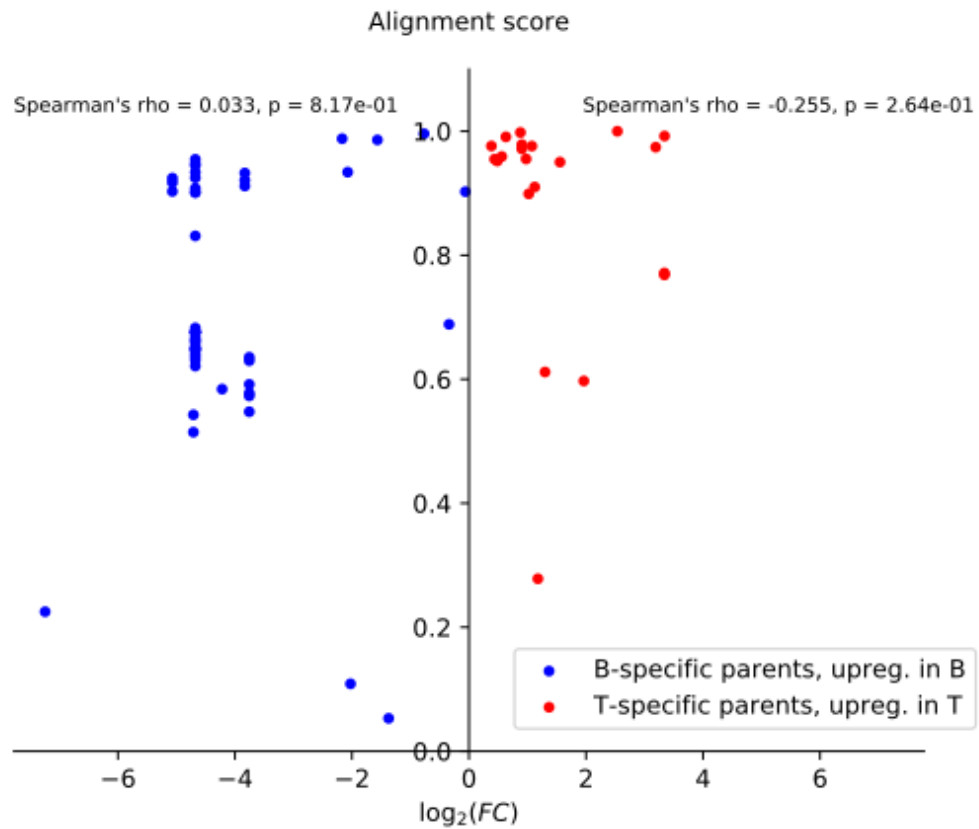


Figure 6.10: A scatter plot showing the log fold change (FC) against the alignment score between retrocopy and parent, for retrocopy parents with a retrocopy expressed in either B cells or T cells. Spearman's rho values do not show strong correlations in either cell type. The bias in expressed retrocopies towards high sequence similarity will skew these results, however.

I carried out a similar analysis comparing the liver samples with the lymphocyte

samples (Figure 6.11). Due to the small number of liver-specific parent transcripts, it is difficult to judge whether there is any bias in their differential expression. However, the lymphocyte-specific parents show the same bias as observed in the B/T cell comparison; namely, upregulation of the parent in the presence of a retrocopy transcript. Examination of the parent transcripts found in liver and lymphocytes showed upregulation in lymphocytes as well. It may be that the discrepancy in number of samples (2 for liver vs. 12 for lymphocytes) skews the differential expression analysis in favour of the larger sample set. Downsampling of the BLUEPRINT samples could be used to remove this discrepancy.

A gene ontology analysis of parent transcripts following the observed bias did not produce any meaningful results. I manually inspected each list (Table 6.7) but did not see any pattern linking all of the parent transcripts. Five genes from the immunoglobulin kappa variable (IGKV) cluster are upregulated in B cells. However, this is to be expected, as these genes are antigen recognition molecules of B cells [252]. In fact, 125 out of 195 IGKV transcripts show upregulation in B cells compared to T cells (Figure 6.12). The remaining 70 do not have sufficiently high confidence values to reliably call differential expression.

Based on this analysis, there is not sufficient evidence to reject the null hypothesis that retrocopy RNA does not affect parent expression. A minority of transcripts with expressed retrocopies show upregulation in the presence of a retrocopy transcript, but there is not evidence to suggest that this is a universal phenomenon. It may be that these few transcripts are indeed regulated by their retrocopies, but if so the mechanism is not clear. Some may be based on sequence identity, but based on the alignment scores this is not necessarily the case. It may be that the retrocopy RNAs act through another mechanism; however, in this case, it is

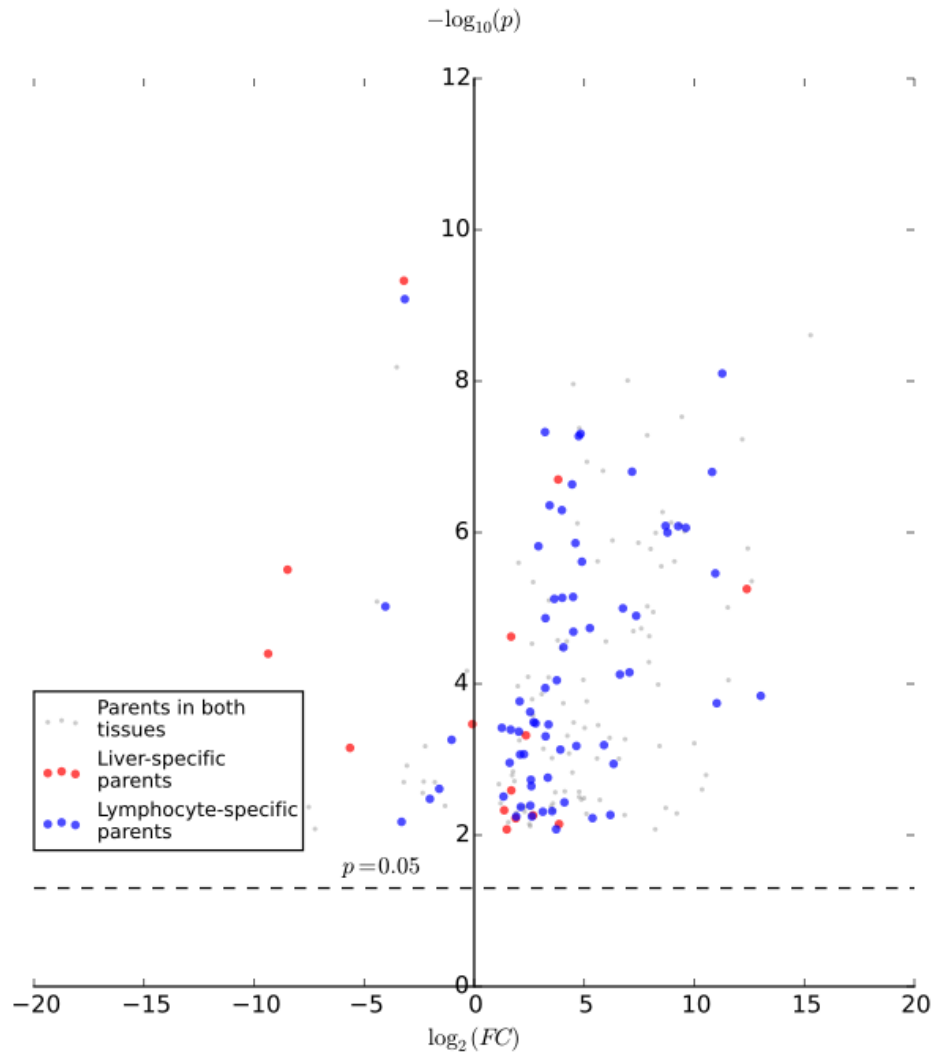


Figure 6.11: Fold change (FC) of retrocopy parents with a retrocopy expressed in either liver or lymphocytes. Positive log FC values indicate upregulation in lymphocytes compared to liver. The small number of liver-specific parents make the results less clear. Examining the shared parents, it appears that there is a general upregulation in lymphocytes.

unclear how or why they should affect the expression of their parent transcripts.

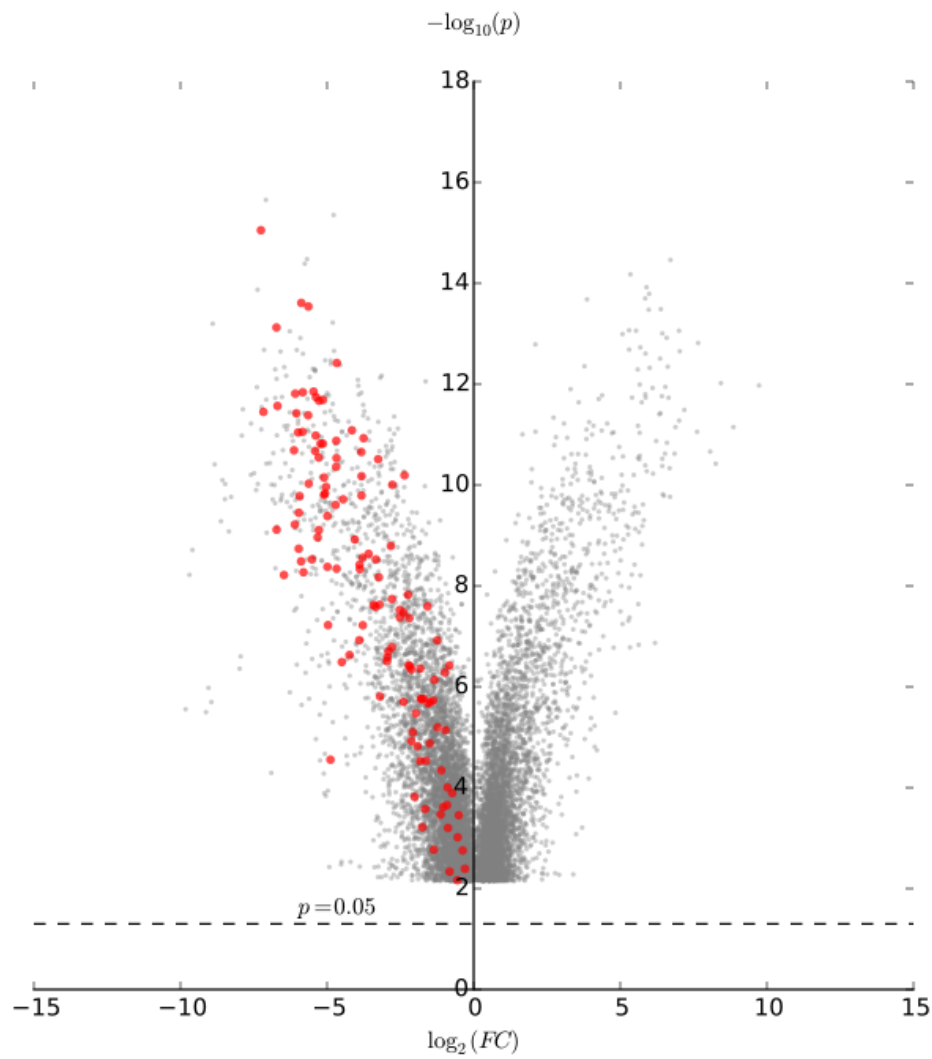


Figure 6.12: Fold change (FC) between B and T values. Immunoglobulin kappa variable transcripts are highlighted in red. Negative log FC indicates upregulation in B cells compared to T cells.

T cell		B cell	
Transcript ID	Gene Name	Transcript ID	Gene Name
ENSMUST00000022925	Eif3h	ENSMUST00000000356	Dazap2
ENSMUST00000022960	Eif3e	ENSMUST000000002436	Snx9
ENSMUST00000025357	Ap3s1	ENSMUST00000029515	S100a11
ENSMUST00000026537	Paox	ENSMUST00000043813	Nudt15
ENSMUST00000028522	Itga6	ENSMUST00000049821	Gm21411
ENSMUST00000031977	Agk	ENSMUST00000058714	Cd24a
ENSMUST00000034983	Atp1b3	ENSMUST00000080204	Sp140
ENSMUST00000035983	Rpl21	ENSMUST00000102795	Ublcp1
ENSMUST00000041048	Orai2	ENSMUST00000115672	Birc3
ENSMUST00000048010	Dse	ENSMUST00000187641	Ncf2
ENSMUST00000051620	Cyb5d1	ENSMUST00000103336	Igkv1-88
ENSMUST00000070215	Npm3	ENSMUST00000103400	Igkv3-5
ENSMUST00000082223	Rpl5	ENSMUST00000116380	Igkv4-53
ENSMUST00000085519	Anp32a	ENSMUST00000103356	Igkv4-57-1
ENSMUST00000097420	Rnaset2a	ENSMUST00000103350	Igkv4-68
ENSMUST00000102840	Ass1		
ENSMUST00000103664	Trav5-4		
ENSMUST00000113064	Traf1		
ENSMUST00000182636	Pdlim1		

Table 6.7: Parent transcripts with at least one retrocopy expressed cell type-specifically, and with upregulation in the presence of retrocopy expression.

6.7 Retrocopy Transcription Does Not Affect Epigenetic State of the Parent Locus

Some studies have found that retrocopy transcripts, and lncRNAs in general, can target epigenetic marks to genes [162,163]. I investigated whether this is the case for cell-specific parents that show differential expression in the direction of the bias described above, i.e., higher expression in the presence of a retrocopy. If rcRNA is targeting epigenetic marks to its parent, we might expect to see regions in or around the parent with cell-type-specific epigenetic state, e.g., a differentially methylated region (DMR).

I first examined the methylation state of each parent transcript, plus 1kb up- and downstream (Datasets and Methods). While some did show DMRs between cell types, I did not see a consistent pattern across all of them, and so there is no evidence of retrocopy transcripts targeting methylation to their parent transcripts. I also looked for differences in histone modification peaks, but similarly, I did not observe any consistent pattern across the parent transcripts (see Online Resources for figures).

Since the transcripts in question are differentially expressed between cell types, different epigenetic states would be expected, and so any differences would not be absolute proof that rcRNAs are having an effect.

6.8 Retrocopy Expression May Affect Protein Abundance

The above analysis on mRNA levels does not provide strong evidence for the regulation of parent transcripts by retrocopy RNA, with a minority of cell-specific parents showing differential expression between cell types. However, if the mechanism is post-transcriptional, this would not necessarily be accompanied by a change in mRNA expression levels, but could be observed in protein abundances. Based on the mRNA analysis, I would expect to see increased parent protein abundance levels in the presence of retrocopy transcripts. In order to investigate this, I used normalised protein abundances from BL6 mouse B and T cells, provided by the Prabakaran group (see Datasets and Methods).

I examined the distribution of protein abundances for parent transcripts with and without an expressed retrocopy (Figure 6.13). Visual inspection suggests that parent transcripts with an expressed retrocopy had protein abundances distributed towards higher values, and with fewer low abundance proteins, compared to parents without an expressed retrocopy, and compared to all proteins. To test this, I used the Anderson-Darling test [253]. For each sex/cell type combination, I compared the abundance distribution for proteins with an expressed retrocopy to those with an unexpressed retrocopy, and to all proteins. A significant result would allow us to reject the null hypothesis that the two sets of abundances were drawn from the same distribution. The results are shown in Table 6.8.

In this case, all of the test statistics exceed the critical value, and so there is sufficient evidence in every case to reject the null hypothesis. We can therefore conclude that proteins translated from transcripts with expressed retrocopies do

tend to have higher abundances compared to other proteins. This is consistent with the observations above that for some transcripts there is upregulation in the presence of an expressed retrocopy.

It is also worth noting that Figure 6.13 appears to show a small difference in abundance distributions between all proteins and those with a retrocopy that is not expressed. This may be because retrocopies are more likely to be formed from genes that are highly and ubiquitously expressed, and this is reflected in protein abundances when we select those that have retrocopies.

Dataset	Number of proteins			Anderson-Darling test		
	expr	unexpr	all	Test pair	A-D statistic	Critical value
Male B	187	889	1932	expr vs. unexpr	18.44	3.75
				expr vs. all	61.27	3.75
Male T	199	888	1932	expr vs. unexpr	15.12	3.75
				expr vs. all	56.11	3.75
Female B	179	892	1932	expr vs. unexpr	8.66	3.75
				expr vs. all	32.71	3.75
Female T	198	888	1932	expr vs. unexpr	23.59	3.75
				expr vs. all	71.16	3.75

Table 6.8: The results of Anderson-Darling (A-D) tests comparing protein abundance distributions (Figure 6.13). The "critical value" represents the value of the A-D statistic with a 1% significance level.

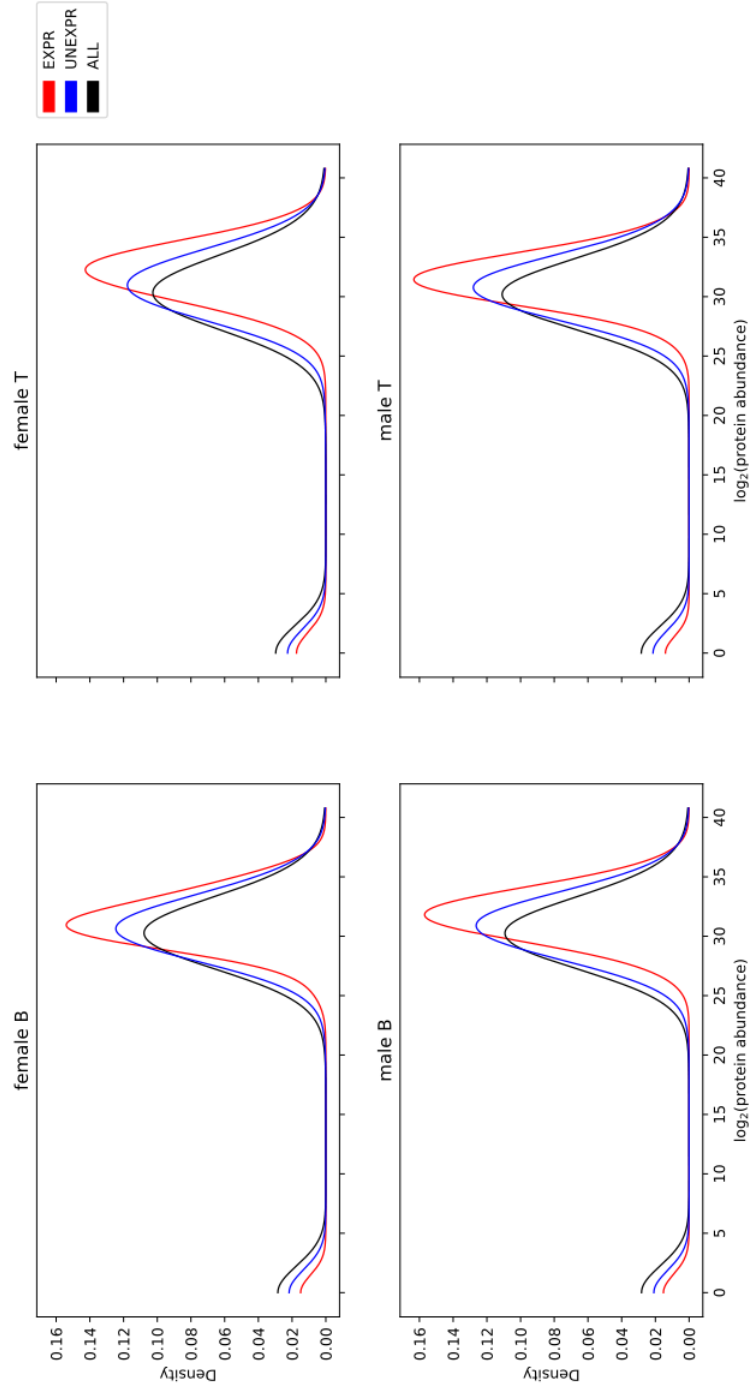


Figure 6.13: Distributions of normalised protein abundance levels for different sets of proteins. EXPR: Proteins with an expressed retrocopy. UNEXPR: Proteins with a retrocopy that is not expressed. ALL: All proteins for which data is available.

6.9 Conservation of BL6 Expressed Retrocopies in CAST

Conservation of genomic features across divergent species or strains implies selection for those features, which can be a sign of functional importance. I therefore assessed the conservation of BL6 retrocopies in CAST, in order to quantify the degree to which those expressed in BL6 are conserved and expressed in CAST.

In order to do this, I first identified which of the BL6 retrocopies are conserved in CAST. After searching for CAST sequences matching BL6 retrocopies, I applied filters to remove false positive and partial hits (see Methods). This left 11,745 putative conserved retrocopies in CAST, across 10,014 BL6 retrocopies.

Table 6.9 shows the results of a chi-squared contingency test between retrocopy expression in BL6 and conservation in CAST. The results suggest that there is a relationship between lack of expression in BL6 and conservation in CAST, the opposite of what one might expect if the retrocopy transcripts have an important function. However, there is a different pattern for retrocopy parents. The results of a chi-squared test are shown in Table 6.10, but in this case the counts are of retrocopy parent transcripts. Here we see that there is a relationship between expression in BL6 and conservation in CAST, so that parents with an expressed retrocopy in BL6 are more likely to have a conserved retrocopy in CAST.

Using the CAST RNA-seq data, I investigated whether retrocopy expression in BL6 was related to expression of conserved retrocopies in CAST. For this analysis, I applied the same RNA-seq analysis pipeline to the CAST data as to the BL6 (see Methods). This produced merged reconstructed transcripts in CAST. I compared these to the conserved retrocopies to produce a list of conserved retrocopies

	CAST Conserved	Not CAST conserved	
BL6 Expressed	338	656	
Not BL6 Expressed	9,676	7,786	
Total BL6 retrocopies			18,456

$$\chi^2 = 172.80$$

$$p = 1.817 \times 10^{-39}$$

Expected Values:

	CAST Conserved	Not CAST conserved	
BL6 Expressed	539.33	454.67	
Not BL6 Expressed	9,474.67	7,987.33	

Table 6.9: The results of a chi-squared contingency test comparing expression of retrocopies in BL6 and their conservation in CAST. The observed values differ significantly from the expected values, suggesting that fewer BL6 expressed retrocopies are conserved than would be expected by chance. However, the significance level of this test is not very low, and the difference between the observed and expected values is not large, so this is not a strong result.

expressed in CAST. I considered only retrocopy parents here, as there are difficulties in establishing exactly which retrocopies are conserved between strains (see Methods). Table 6.11 summarises the results of a chi-squared contingency test. There is a significant relationship between expression in BL6 and conservation and expression in CAST.

Intuitively, it would not be surprising if retrocopies expressed in BL6 and conserved in CAST were also expressed in CAST. However, this analysis indicates that there is not a relationship between expression of a retrocopy in BL6 and its

	Has CAST Conserved Retrocopy	No CAST Conserved Retrocopy	
BL6 Expressed	351	105	
Not BL6 Expressed	2,315	1,089	
Total BL6 retrocopy parents			3,860

$$\chi^2 = 14.71$$

$$p = 1.25 \times 10^{-4}$$

Expected Values:

	Has CAST Conserved Retrocopy	No CAST Conserved Retrocopy	
BL6 Expressed	314.95	141.05	
Not BL6 Expressed	2,351.05	1,052.95	

Table 6.10: The results of a chi-squared contingency test comparing parents of expressed retrocopies in BL6 and the parents of retrocopies conserved in CAST. The observed values differ significantly from the expected values, suggesting that there are more parents with both an expressed retrocopy in BL6 and a conserved retrocopy in CAST.

expression in CAST; indeed, it appears that BL6 expressed retrocopies are *less* likely to be conserved in CAST (Table 6.9). Given that expressed retrocopies in BL6 shown signs of being younger, this would make sense, if these specific retrocopies appeared after CAST and BL6 diverged. What these analyses do suggest is that parent genes with a retrocopy expressed in BL6 are more likely to have a retrocopy conserved and expressed in CAST, even if it is not the same retrocopy. It would appear, then, that certain parent genes are more likely to have expressed retrocopies across divergent strains. It may be that there is selection in favour

	CAST expressed	Not CAST expressed	
BL6 Expressed	173	178	
Not BL6 Expressed	160	2,155	
Total BL6 retrocopy parents conserved in CAST			2,666

$$\chi^2 = 496.86$$

$$p = 4.58 \times 10^{-110}$$

Expected Values:

	CAST expressed	Not CAST expressed	
BL6 Expressed	43.84	307.16	
Not BL6 Expressed	289.16	2,025.84	

Table 6.11: The results of a chi-squared contingency test comparing parents of expressed retrocopies in BL6 and the parents of retrocopies conserved and expressed in CAST. The observed values differ significantly from the expected values, suggesting that there are more parents with both an expressed retrocopy in BL6 and a conserved and expressed retrocopy in CAST.

of these parents having expressed retrocopies, and due to the dynamic nature of retrocopy formation, different individual retrocopies are expressed according to precise conditions.

This is consistent with the earlier observations that there is a set of parents with expressed retrocopies shared between B, T, and liver cells. I compared this set of parents to the parents that have conserved and expressed retrocopies in CAST (Table 6.12). While the result is not strongly significant compared to other results in this section, the p value shows significance at the 1% level. This suggests that there is a link between having at least one retrocopy expressed in BL6 across all

three cell types, and having a conserved and expressed retrocopy in CAST. This would again be consistent with the existence of a set of parents for which retrocopy expression is advantageous, possibly because they are regulated by retrocopy RNA. An important extension to this work would be to use a full retrocopy annotation in CAST to confirm this possibility, as the retrocopy discovery method used here is restricted to those that exist in both BL6 and CAST. It may be that these parent genes have CAST-specific retrocopies that are expressed.

	CAST expressed	Not CAST expressed
BL6 Expressed and Shared	95	76
BL6 Expressed, not Shared	92	133
Total BL6 cell-shared retrocopy parents conserved in CAST		396

$$\chi^2 = 7.81$$

$$p = 5.20 \times 10^{-3}$$

Expected Values:

	CAST expressed	Not CAST expressed
BL6 Expressed and Shared	80.75	90.25
BL6 Expressed, not Shared	106.25	118.75

Table 6.12: The results of a chi-squared contingency test comparing cell-shared parents of expressed retrocopies in BL6 and the parents of retrocopies conserved and expressed in CAST. That is, a retrocopy parent which has at least one retrocopy expressed in all three BL6 cell types will fall into the “BL6 Expressed and Shared” category; a parent with a retrocopy expressed in only one or two cell types will fall into the “BL6 Expressed, not Shared” category. The observed values differ significantly from the expected values, suggesting that there are more parents with both an expressed retrocopy in BL6 across B, T, and liver cells, and a conserved and expressed retrocopy in CAST.

6.10 Summary

In this chapter, I have demonstrated a significant bias towards retrocopy RNA complementary to the parent transcript RNA and with high sequence identity to the parent. These retrocopy RNAs therefore have the potential to regulate their

parent transcripts through the formation of RNA:RNA duplexes. While there is not strong evidence for this at the mRNA level, protein abundance values are higher for transcripts with an expressed retrocopy, suggesting that the retrocopy RNAs may play a protective role for their parent transcripts. By comparing cell types and mouse strains, I have identified sets of parent transcripts that tend to have expressed retrocopies across cell types and lineages. Specific retrocopies are not necessarily expressed in all cases, and do not show conservation across mouse strains. I propose that there are in fact a set of protein-coding genes that are post-transcriptionally regulated by retrocopies. Due to the risk of decay of retrocopies and the different chromatin environments in different cell types, these parents may be regulated by different specific retrocopies in different circumstances. This would be a novel regulatory mechanism that ensures consistent and reliable regulation by taking advantage of the redundancy introduced by multiple retrocopies.

Chapter 7

Discussion

7.1 Retrotransposon Transcription

7.1.1 Summary of Results

In this thesis, I have quantified the retrotransposon content of whole transcriptomes in mouse B and T lymphocytes. I found that 40-45% of transcripts contain some retrotransposon content, but the majority of these have relatively low retrotransposon content. A small proportion (3-5%) have high retrotransposon content (>50%), and these can be separated into distinct clusters based on their retrotransposon content. There are particularly distinct clusters of transcripts consisting entirely of L1 or ERV sequence. Similar clusters are found in both B and T cells, as well as liver cells; however, the individual retrotransposon elements being transcribed are not necessarily the same, and show cell-type specificity. I also found a potential regulatory role for retrotransposons in ncRNAs: antisense transcript expression is more strongly correlated with expression of the corresponding sense

transcript if the antisense transcript contains retrotransposon sequence, and particularly if the sense/antisense pair both contain the same type of retrotransposon. This indicates a possible role for retrotransposons in regulation of transcripts by their antisense counterpart.

7.1.2 Comparison to Existing Literature

As discussed in the Introduction, there are relatively few studies that have quantified the retrotransposon content of somatic cell transcriptomes; of these, only a small number use a mouse model. In particular, two studies have carried out transcriptome-scale analysis of retrotransposons in mouse transcripts, although both focused on lncRNAs. Kapusta *et al.* used available lncRNA annotations in human, mouse, and zebrafish and compared them to the RepeatMasker annotation in each species [128]. The mouse data comprised 2,167 transcripts. Kannan *et al.* used microarray data for long intergenic ncRNAs (lincRNAs) in mouse [254], downloaded from NRED [255], and compared these to RepeatMasker. This dataset comprised 2,390 transcripts. In both cases, multiple cell types and tissues were included.

These studies reported 68% (Kapusta) and 51% (Kannan) of mouse lncRNAs containing transposable element (TE) sequence. The lower value found in the Kannan study may be due to the use of microarray data, and to the inclusion of only intergenic lncRNAs. Both studies report a higher proportion of transcripts with TE content than I do; however, my estimates include all transcripts, rather than only lncRNAs, and so this is expected. While I do identify putative lncRNAs based on comparison with protein-coding genes, it would not be meaningful to

analyse these as lncRNAs without further validation that they can be classified as such; for example, by an open reading frame analysis.

There is clearer agreement between both previous studies and my own work when examining the proportion of each transcript comprising TEs. Kapusta *et al.* reported that in humans the majority of lncRNAs have relatively low TE content: approximately 20% have more than 50% TE sequence, and around 1% have more than 90% TE sequence. However, they did not report an equivalent analysis in mouse. Kannan *et al.* report that 78% of mouse lncRNAs have less than 20% TE sequence; supplementary figures show around 1% of lncRNAs have more than 50% TE sequence, and almost none with more than 90%. This is broadly in agreement with my findings, in which 95% of mouse transcripts have less than 50% retrotransposon sequence. The differences between exact figures may again be due to the differences in datasets and methods used; in particular, the exclusive use of lncRNAs by Kannan *et al.* Comparisons with the percentages from Kapusta *et al.* can only be made in the broadest terms, as these figures are for human.

There is less clear agreement on the exact types of TE included. In the mouse ENSEMBL lncRNAs, the Kapusta study reported that, as a percentage of total sequence, LTRs contributed 40%, SINEs 30%, and LINEs 25%, with other non-retrotransposon TEs contributing the remaining 5% (mainly DNA transposons). Kannan *et al.* performed a similar analysis. Although their methods are not directly comparable, the order of contribution is the same: LTRs, SINEs, LINEs, and DNA transposons. Here, I find that approximately 75% of the retrotransposons found inside a transcript are SINEs, with LTRs and LINEs making up 14% and 10%, respectively. The discrepancy between my results and those from Kapusta and colleagues may be due to the relative sizes of the different retrotrans-

poson classes: SINEs tend to be much shorter than LTRs and LINEs, and so a large number may still contribute a relatively small proportion of sequence. The sequence contributions shown in Chapter 5 are more closely in agreement with the previous studies.

Neither of these studies analyse the cell-type specificity of retrotransposon expression; the closest such analysis I found is from Faulkner *et al.*, which focused on retrotransposon-derived transcriptional start sites (TSSs) [127]. This study reported that the activity of different classes of retrotransposon is cell-type specific. This is in direct contradiction to my findings, which is that the same types of retrotransposon are expressed across different cell types, but the individual elements are specific. Broadening the cell types assessed may reveal different results, but the results across the three cell types already included are remarkably consistent. However, the activity of retrotransposon TSSs and actual retrotransposon transcription are different phenomena, and comparisons should be made with caution.

The regulation of mRNAs by their corresponding antisense lncRNA is well-studied (reviewed in [246–248]). At the transcriptional level, lncRNA transcription can recruit chromatin modifying complexes, thus influencing expression *in cis* [246]. At the post-transcriptional level, the formation of mRNA:lncRNA duplexes can either up- or down-regulate the target mRNA, depending on the context. For example, this can mask miRNA binding sites [247], or influence splicing [248]. RNA:RNA duplexes can also be targeted by enzymes. For example, stau1 (STAU1) mediated decay (SMD) degrades mRNA when STAU1 binds to double stranded RNA [256]. Several studies have reported that the formation of these duplexes is dependent on the presence of a SINE in the UTR of the mRNA, to which a lncRNA containing a complementary SINE binds, in multiple species [256, 257].

This demonstrates the potential role for retrotransposons in RNA:RNA binding, but is a single example, involving only a single class of retrotransposons. The evidence presented here suggests that a similar phenomenon may be part of the regulatory pathway of many genes, using different classes of retrotransposon. Johnson and Guigó have previously suggested that retrotransposons act as functional domains in lncRNAs, and in particular as sites for hybridisation to other RNA and DNA molecules [130]. Such a role is consistent with the observations reported here; however, it is unclear how retrotransposon sequence would present an advantage over other sense/antisense pairs. In addition, the mechanism through which any such regulation occurs is unknown. The fact that the bias is seen at the mRNA level suggests that it occurs at the transcriptional or post-transcriptional stage, rather than at the translation stage. The bias towards up-regulation suggests that if the mechanism is post-transcriptional then it works through a protective mechanism, such as miRNA site masking.

7.1.3 Conclusions and Future Work

As expected, retrotransposons make significant contributions to the transcriptomes of somatic cells. Transcripts can be classified based on the amount of retrotransposon sequence they contain, and the type of retrotransposons included. In particular, there are relatively small but distinct sets of transcripts consisting almost entirely of ERV and LINE sequence. These clusters are found in diverse cell types, but the individual retrotransposon elements are cell type-specific. If these are functional, it may be that cells utilise different retrotransposon elements to fulfil the same role under different epigenetic conditions.

It is possible, however, that these transcripts instead represent transcriptional noise, or possibly retrotransposition intermediates. To rule out these latter possibilities, future work should include the following:

- Repeat analysis with different parameters and annotations to ensure results are not an artefact of parameter choice
- Expand this analysis to other cell types to ensure that these clusters are consistent
- Expand this analysis to other mouse strains (beginning with CAST), as these transcripts are unlikely to be BL6-specific if they represent true functional transcripts

Similar results from all three extensions would confirm that these transcripts are not technical noise. A more detailed study of their exact retrotransposon content would also be useful. For example, are they full-length retrotransposons?

The next step would be to test for function, ideally using experimental assays rather than bioinformatic techniques. Knockout or knockdown experiments would be ideal, but this may be difficult given the number of transcripts involved, and their repetitive nature, which would increase the likelihood of off-target effects (which could be severely deleterious given the contribution of retrotransposons to regulatory regions). Liu *et al.* recently demonstrated the use of a modified CRISPR assay to inhibit expression of multiple lncRNAs simultaneously [25]; a similar technique could be useful here. Bioinformatics analysis of co-expression between these transcripts and coding transcripts could be used to select candidates.

If such studies successfully showed a function for these transcripts, this would be evidence for distinct classes of lncRNA common to many somatic cell types,

with the high retrotransposon content currently associated with transcripts found in pluripotent cells [115–118, 120–126]. It is too early to speculate what these functions might be, but the well-established contributions of retrotransposons to regulatory networks is promising.

The widespread role of retrotransposons in sense/antisense pairs is also novel, and suggests that retrotransposons facilitate up-regulation or protection of complementary transcripts. While specific examples of retrotransposons in this kind of regulation have been identified, this is the first evidence for such action on a broad scale. However, these findings are currently only statistical, and must be validated. If the three tasks listed above demonstrated similar results, this would be a promising start, reducing the likelihood that this is a technical artefact. Experimental validation of these results and establishing the mechanism would also be important. Experimental validation is made difficult by the fact that sense/antisense pairs cannot be knocked out separately. Bioinformatics analysis could identify putative promoters for the antisense transcripts, which could then be used to inhibit expression of the antisense transcript. The CRISPRi technique mentioned above [25] could also be useful for assaying multiple lncRNAs.

To discover the mechanism, a relatively simple first step would be to examine the shared retrotransposons between the sense and antisense transcripts. What kind of retrotransposons are represented? How large are they? Which part of the transcript are they located in? Are they always the same retrotransposon element? Bioinformatics tools have been developed to computationally predict RNA:RNA interactions [258–260]. The results of such a tool could be compared to the results from this work; if there is evidence for the formation of these duplexes, they could be validated using experimental techniques. If not, exploration of transcriptional

mechanisms would be warranted.

7.2 Retrocopy Transcription

7.2.1 Summary of Results

In this thesis, I have found that retrocopies are expressed throughout the genome in B and T lymphocytes and liver cells, originating from a relatively restricted pool of parent transcripts. I have found a small number of retrocopy/parent pairs showing evidence for up-regulation of the parent mRNA when a retrocopy is expressed. I also found evidence for increased abundance levels of proteins translated from parent transcripts when a matching retrocopy is expressed, suggesting a possible regulatory role for retrocopy transcripts at the post-transcriptional or translational level. I also found that expressed retrocopy RNA is almost always complementary to the parent mRNA, and tends to have high sequence identity to the parent mRNA. This would allow for a regulatory mechanism based on the formation of RNA:RNA duplexes.

I compared the retrocopy transcripts in different cell types, and found that there is enrichment for parent transcripts with expressed retrocopies in multiple cell types, although the individual retrocopies may be different in each cell type. When comparing BL6 and CAST mouse strains, I found that while specific retrocopies are not necessarily conserved between strains, there is enrichment for parent transcripts with expressed retrocopies in both strains. This suggests that there may be a set of parent transcripts that rely on retrocopy-based regulation, but utilise different retrocopies depending on the genomic or epigenetic context.

7.2.2 Comparison to Existing Literature

Several previous studies have also quantified retrocopy transcription in mouse, as well as in human and other species. The following three studies report comparable results to the work presented here:

- Yano *et al.* [261]: by comparing expressed sequence tags (ESTs) to a database of 4,476 previously identified mouse retrocopies, Yano and colleagues found evidence for expression of between 22 and 45 retrocopies (0.5-1%)
- Harrison *et al.* [262]: using a novel method, Harrison and colleagues identified 215 transcribed retrocopies (0.04%) in mouse by comparing annotated genes and ESTs to a previously published set of 5,582 retrocopies [144]
- Carelli *et al.* [141]: using a novel method, Carelli and colleagues identified 5,569 retrocopies in mouse, of which 420 (7.5%) show signs of expression (FPKM >1), across 6 tissues

In this work, I report 1,131 expressed retrocopies across three cell types, out of a total of 18,456 retrocopies (6.12%). This is a comparable proportion to that reported in the Carelli study, and significantly more than the proportion found to be expressed by either the Yano or Harrison studies. Both the Yano and Harrison studies are based on ESTs, and data from more than a decade ago, while this work and the Carelli study are based on RNA-seq, and more recent data. These two factors may account for the discrepancy in the proportion of expressed retrocopies. ESTs may not be effective at detecting low abundance transcripts [263], although I was unable to find a direct comparison between the ESTs and RNA-seq. None of these studies contained a detailed analysis of the parent genes from which the

retrocopies originated, or a comparison between the retrocopies and their parents.

The idea that retrocopy antisense transcripts might influence the expression of their parent transcripts is not new, and was first theorised in 1986 by McCarrey and Riggs [264]. The first example of this to be characterised was by Korneev *et al.* in a snail model [172], who showed that the antisense transcript from a nitric oxide synthase (NOS) retrocopy suppresses NOS, although the exact mechanism was not established. More recently, Hawkins and Morris showed that Oct4 is suppressed by an antisense Oct4 retrocopy transcript in human cell lines [265], and suggested a mechanism where the retrocopy RNA targets epigenetic regulators to the parent gene. In 2013, Johnsson *et al.* found a similar example in human cells: the phosphatase and tensin homolog (PTEN) gene, which has a corresponding antisense retrocopy transcript with two isoforms. One isoform suppresses the parent transcript by targeting epigenetic modulators to PTEN, while the other forms an RNA:RNA duplex with the PTEN mRNA, reducing translation. It should be noted that in all cases the antisense retrocopy transcript has a repressive effect on the parent. In addition to these specific examples, Muro and Andrade-Navarro identified 87 such transcripts using ESTs in a genome-wide screen [159], but did not investigate whether these RNAs interacted with the parent transcripts.

The work I present here represents the first transcriptome level analysis of retrocopy regulation of parent transcripts. My results suggest that the majority of these interactions are based on sequence complementarity, as observed in the examples described above. In addition to these retrocopy-based examples, there is a significant body of literature describing the interactions between sense/antisense transcript pairs, as discussed above. However, unlike the examples of NOS, Oct4, and PTEN, my results suggest that the presence of an antisense retrocopy tran-

script is linked to increased protein abundance, rather than repression. While this has not previously been observed in retrocopy/parent interactions, there is evidence for this in the general sense/antisense pair regulation literature [247]. Possible mechanisms that would explain this include the formation of RNA:RNA duplexes that mask miRNA binding sites, thus preventing miRNA-based degradation of the parent transcript. The antisense transcripts could recruit activating epigenetic marks to the parent locus, although in this case a clearer up-regulation of mRNA would be expected, whereas here the only clear increase is in protein abundance.

My findings on cell type specificity and conservation between strains also represent a departure from existing literature. Carelli *et al.* compared cell types and species, but focused on individual retrocopies. Carelli and colleagues showed that in mouse, retrocopy transcription tends to be cell type specific. Here, I have shown that while expression of specific retrocopies shows some cell type specificity, the same parents tend to have expressed retrocopies across cell types, and a similar trend is seen when comparing strains. While more work is needed to validate these findings, I believe it may indicate that the parent transcripts are the more important consideration, especially given the potential regulatory effects of expressed retrocopies.

7.2.3 Conclusions and Future Work

Based on my results, I hypothesise the existence of a novel regulatory mechanism based on complementary RNAs transcribed from retrocopies. These transcripts positively regulate their parent transcripts at the post-transcriptional level, either

through a protective mechanism that prevents degradation, or an active mechanism that encourages translation. Given the high level of sequence similarity observed in expressed retrocopies, the mechanism could be based on the formation of RNA:RNA duplexes. Further, I would suggest that the protein-coding genes that are regulated in this manner may utilise different retrocopy instances under different circumstances to ensure consistent regulation, regardless of genetic or epigenetic context.

There are some clear parallels between the work presented on antisense transcripts with retrotransposons, and the antisense retrocopy transcripts. It would be worth investigating whether there is retrotransposon sequence common to both the parent transcript and the retrocopy transcript, and whether this sequence is linked to stronger upregulation, as seen in Chapter 5.

I believe that the results presented in this thesis are a compelling basis for further investigation. The initial steps would be similar to those described for my results on retrotransposons:

- Repeat analysis with different parameters and annotations to ensure results are not an artefact of parameter choice
- Expand this analysis to other cell types
- Recreate full retrocopy expression analysis in CAST
- Expand this analysis to other mouse strains and species

The first point is particularly important given the number and diversity of retrocopy annotations available. I would hope to see essentially the same retrocopies and parents being expressed, regardless of the annotation used, rather than being

an artefact of the retrocopy annotation I used. The second and third points are essential in establishing whether there is a set of parent genes that consistently have expressed retrocopies, regardless of context.

This work has not established how the expressed retrocopies are regulated. While I was able to find interesting links between the retrocopy transcripts and flanking retrotransposon sequence, I could not establish whether the retrotransposons were functioning as promoters. A bioinformatics analysis could be used to identify putative promoters for the expressed retrocopies, potentially followed by experimental validation.

Experimental validation would also be essential in establishing whether retrocopies do actually regulate their parents, and the mechanism of action if this is the case. Knockout studies on a set of candidate parent/retrocopy pairs would establish whether removing the retrocopy affects the mRNA level or protein abundance of the parent. In addition, the CRISPRi technique mentioned previously could also be used to target multiple retrocopies simultaneously [25]. If these experiments confirmed the link between retrocopy expression and parent expression, the next step would be to establish a mechanism. Bioinformatic analysis could be used to establish whether RNA:RNA duplex formation would be possible; if so, experimental validation could follow. Known databases of miRNA targets could be used to check for miRNA binding sites in parent transcripts that could be masked by an antisense retrocopy RNA. This is an ambitious program of work, but, if successful, it would produce compelling evidence for an important new regulatory mechanism.

7.3 Final Remarks

The findings I have presented should be treated as what they are: the results of a purely bioinformatic analysis of high-throughput sequencing (HTS) data. While RNA-seq and other HTS techniques are a powerful tool for exploring the genome, transcriptome, and epigenome, they are not flawless, and should be treated with a degree of caution. The identification of lncRNAs, for example, is fraught with possible sources of error. RNA-seq itself has intrinsic biases, aligners can map reads incorrectly, and reconstruction of transcriptomes can produce significant numbers of false positive transcripts. All of these may be compounded, on top of biological noise, to produce misleading results. High-quality software tools and well-designed bioinformatics pipelines can serve to ameliorate these errors, but even then a different choice of software can give rise to notably different results. It is therefore sensible to allow for the possibility that the results of a bioinformatics pipeline will contain a certain degree of noise. Nonetheless, these analyses are useful, and can be used to estimate the number of transcripts and their sequence content. Experimental validation is essential.

Biological noise should also be considered. The widely publicised findings from ENCODE and similar large-scale projects claim that almost the entire genome is transcribed, with the implication that every one of these transcripts is a useful, functional molecule. This could well be the case, or it may be that in fact many of the observed transcripts are simply “genomic weather”, the result of accidental transcriptional activity arising from essentially stochastic processes and the presence of cryptic promoters. It would be ignorant to assume that every non-coding transcript is meaningless, and increasing numbers of non-coding RNAs are being

functionally characterised. However, given the number of ncRNAs reported, how likely is it that every single one has a specific function? Or that they are all involved in some kind of coordinate action, as proposed by Melé and Rinn [18]? To me, the latter seems more likely, combined with a certain amount of biological noise.

Ideally, one would experimentally validate every single RNA molecule that is functionally characterised, but this is far from practical given that the number of reported molecules is often on the order of 10,000. The validation of a few candidate molecules is of limited use: it is seldom possible to do this for more than a small percentage of the molecules found, and functional characterisation is even more difficult. In addition, lncRNAs present their own set of technical problems associated with experimental validation, and are not always amenable to the techniques developed for coding genes [266].

There is hope on the horizon though. Classical genetics infers function based on the effects of removal, and the advent of CRISPR-based techniques for genome editing are making it easier to accurately remove or insert sequence of interest. The effects of mass retrotransposon deletion would certainly be interesting. From a bioinformatics perspective, the ever-decreasing cost of short-read sequencing and the increasing length of short reads means that more accurate results can be obtained with increased depth and mapping precision. Bioinformatics software is ever-improving. The emergence of long-read technologies promises much-improved mapping, with the promise of banishing the problem of multimapping reads, which will be particularly useful for the study of repetitive regions, and researchers have already started to do so [129]. Single-cell sequencing could be used to establish whether transcripts are expressed consistently across individual cells in a popu-

lation. All together, these emerging technologies promise to vastly improve the accuracy with which we can measure transcription and related phenomena, leading to a better understanding of how the vast catalogue of RNA helps to shape us.

Bibliography

- [1] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids. *Nature*, 1953.
- [2] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 1961.
- [3] S. Brenner, F. Jacob, and M. Meselson. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 1961.
- [4] F. Gros *et al.* Unstable ribonucleic acid revealed by pulse labelling of *Escherichia coli*. *Nature*, 1961.
- [5] M. Cobb. Who discovered messenger RNA? *Current Biology*, 2015.
- [6] B. McClintock. The origin and behavior of mutable loci in maize. *PNAS*, 1950.
- [7] N. C. Comfort. "The Real Point Is Control": The Reception of Barbara McClintock's Controlling Elements. *Journal of the History of Biology*, 1999.
- [8] A. F. A. Smit, R. Hubley, and P. Green. RepeatMasker Open-4.0., 2013-2015.
- [9] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, 5 edition, 2008.
- [10] T. R. Cech and J. A. Steitz. The Noncoding RNA Revolution: Trashing Old Rules to Forge New Ones. *Cell*, 2014.
- [11] K. W. Vance and C. P. Ponting. Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends in Genetics*, 2014.
- [12] K. F. Tóth, D. Pezic, E. Stuwe, and A. Webster. The piRNA Pathway Guards the Germline Genome Against Transposable Elements. In D. Wilhelm and P. Bernard, editors, *Non-coding RNA and the Reproductive System*, chapter 4. Springer Science+Business Media, 2016.
- [13] M. N. Cabili *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes and Development*, 2011.
- [14] T. Derrien *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research*, 2012.
- [15] M. J. Hangauer *et al.* Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS Genetics*, 2013.
- [16] P. Carninci *et al.* The transcriptional landscape of the mammalian genome. *Science*, 2005.
- [17] C.-C. Hon *et al.* An atlas of human long non-coding RNAs with accurate 5' ends. *Nature*, 2017.
- [18] M. Melé and J. L. Rinn. "Cat's Cradling" the 3D Genome by the Act of LncRNA Transcription. *Molecular Cell Perspective*, 2016.

- [19] H. Jia *et al.* Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA*, 2010.
- [20] J. Harrow *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*, 2012.
- [21] M. K. Iyer *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics*, 2015.
- [22] Y. Zhao *et al.* NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Research*, 2016.
- [23] J. M. Mudge, A. Frankish, and J. Harrow. Functional transcriptomics in the post-ENCODE era. *Genome Research*, 2013.
- [24] M. Joaquina Delàs *et al.* lncRNA requirements for mouse acute myeloid leukemia and normal differentiation. *eLife*, 2017.
- [25] S. J. Liu *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science*, 2017.
- [26] M. Safran *et al.* GeneCards Version 3: the human gene integrator. *Database*, 2010. www.genecards.org.
- [27] R. Galupa and E. Heard. X-chromosome inactivation: new insights into cis and trans regulation. *Current Opinion in Genetics and Development*, 2015.
- [28] E. Hacisuleyman *et al.* Topological organization of multichromosomal regions by the long intergenic non-coding RNA Firre. *Nature Structural and Molecular Biology*, 2014.
- [29] P. Schorderet and D. Duboule. Structural and Functional Differences in the Long Non-Coding RNA Hotair in Mouse and Human. *PLoS Genetics*, 2011.
- [30] A. R. Amândio *et al.* Hotair Is Dispensable for Mouse Development. *PLoS Genetics*, 2016.
- [31] N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, editors. *Mobile DNA II*. American Society for Microbiology, 2002.
- [32] S. T. Szak *et al.* Molecular archeology of L1 insertions in the human genome. *Genome Biology*, 2002.
- [33] S. L. Martin *et al.* The structures of mouse and human L1 elements reflect their insertion mechanism. *Cytogenetic Genome Research*, 2005.
- [34] B. Brouha *et al.* Hot L1s account for the bulk of retrotransposition in the human population. *PNAS*, 2003.
- [35] N. de Parseval *et al.* Survey of Human Genes of Retroviral Origin: Identification and Transcriptome of the Genes with Coding Capacity for Complete Envelope Proteins. *Journal of Virology*, 2003.
- [36] R. K. Slotkin and R. Martienssen. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, 2007.
- [37] D. J. Finnegan. Retrotransposons. *Current Biology*, 2012.
- [38] S. Boissinot and A. Sookdeo. The Evolution of LINE-1 in Vertebrates. *Genome Biology and Evolution*, 2016.
- [39] C. R. Beck *et al.* LINE-1 Elements in Structural Variation and Disease. *Annual Review of Genomics and Human Genetics*, 2011.
- [40] C. Esnault, J. Maestre, and T. Heidmann. Human LINE retrotransposons generate processed pseudogenes. *Nature Genetics*, 2000.

- [41] D. A. Kulpa and J. V. Moran. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. *Nature Structural and Molecular Biology*, 2006.
- [42] W. Wei *et al.* Human L1 retrotransposition: cis preference versus trans complementation. *Molecular Cell Biology*, 2001.
- [43] E. M. Ostertag and Jr. H. H. Kazazian. Twin Priming: A Proposed Mechanism for the Creation of Inversions in L1 Retrotransposition. *Genome Research*, 2001.
- [44] A. M. Lambowitz and Marlene Belfort. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *American Society for Microbiology*, 2015.
- [45] M. Dewannieux, C. Esnault, and T. Heidmann. LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics*, 2003.
- [46] M. Dewannieux and T. Heidmann. L1-mediated Retrotransposition of Murine B1 and B2 SINEs Recapitulated in Cultured Cells. *Journal of Molecular Biology*, 2005.
- [47] K. Ohshima and N. Okada. SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. *Cytogenetic Genome Research*, 2005.
- [48] S. R. Richardson *et al.* The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiology Spectrum*, 2015.
- [49] G. R. Daniels and P. L. Deininger. Repeat sequence families derived from mammalian tRNA genes. *Nature*, 1985.
- [50] N. A. Vassetzky and D. A. Kramerov. SINEBase: a database and tool for SINE analysis. *Nucleic Acids Research*, 2013.
- [51] A. F. Smit. The origin of interspersed repeats in the human genome. *Current Opinion in Genetics and Development*, 1996.
- [52] D. L. Mager and J. P. Stoye. Mammalian Endogenous Retroviruses. *Microbiology Spectrum*, 2015.
- [53] N. G. Copeland, K. W. Hutchison, and N. A. Jenkins. Excision of the DBA ecotropic provirus in dilute coat-color revertants of mice occurs by homologous recombination involving the viral LTRs. *Cell*, 1983.
- [54] J. M. Coffin, S. H. Hughes, and H. E. Varmus, editors. *Retroviruses*. Cold Spring Harbor Laboratory Press, 1997.
- [55] C. Stocking and C. A. Kozak. Murine endogenous retroviruses. *Cellular and Molecular Life Sciences*, 2008.
- [56] D. C. Hancks and H. H. Kazazian Jr. Active human retrotransposons: variation and disease. *Current Opinion in Genetics and Development*, 2012.
- [57] K. H. Burns. Transposable elements in cancer. *Nature Reviews Cancer*, 2017.
- [58] J. A. Yoder, C. P. Walsh, and T. H. Bestor. Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, 1997.
- [59] By Mariuswalter - Own work, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=54318073>.
- [60] M. G. Goll and T. H. Bestor. Eukaryotic Cytosine Methyltransferases. *Annual Review of Biochemistry*, 2005.
- [61] C. P. Walsh, J. R. Chaillet, and T. H. Bestor. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature Genetics*, 1998.

- [62] D. Bourc’his and T. H. Bestor. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature*, 2004.
- [63] R. Jaenisch, A. Schnieke, and K. Harbers. Treatment of mice with 5-azacytidine efficiently activates silent retroviral genomes in different tissues. *PNAS*, 1985.
- [64] L. Lavie *et al.* CpG methylation directly regulates transcriptional activity of the human endogenous retrovirus family HERV-K(HML-2). *Journal of Virology*, 2005.
- [65] D. Reiss, Y. Zhang, and D. L. Mager. Widely variable endogenous retroviral methylation levels in human placenta. *Nucleic Acids Research*, 2007.
- [66] G. Howard *et al.* Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. *Oncogene*, 2008.
- [67] S. Stengel *et al.* Regulation of human endogenous retrovirus-K expression in melanomas by CpG methylation. *Genes Chromosomes Cancer*, 2010.
- [68] S. Szpakowski *et al.* Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements. *Gene*, 2009.
- [69] D. Varshney *et al.* SINE transcription by RNA polymerase III is suppressed by histone methylation but not by DNA methylation. *Nature Communications*, 2015.
- [70] S. Seisenberger *et al.* Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers. *Philosophical Transactions of the Royal Society B*, 2013.
- [71] J. H. A. Martens *et al.* The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *The EMBO Journal*, 2005.
- [72] T. S. Mikkelsen *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 2007.
- [73] T. Matsui *et al.* Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature*, 2010.
- [74] H. M. Rowe *et al.* KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature*, 2010.
- [75] M. A. Carmell *et al.* MIWI2 Is Essential for Spermatogenesis and Repression of Transposons in the Mouse Male Germline. *Developmental Cell*, 2007.
- [76] W. Deng and H. Lin. *miwi*, a Murine Homolog of *piwi*, Encodes a Cytoplasmic Protein Essential for Spermatogenesis. *Developmental Cell*, 2002.
- [77] S. Kuramochi-Miyagawa *et al.* Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. *Development*, 2004.
- [78] E.-M. Weick and E. A. Miska. piRNAs: from biogenesis to function. *Development*, 2014.
- [79] R. F. Ketting. The Many Faces of RNAi. *Developmental Cell*, 2011.
- [80] E. F. Roovers *et al.* Piwi Proteins and piRNAs in Mammalian Oocytes and Early Embryos. *Cell Reports*, 2015.
- [81] A. A. Aravin *et al.* A piRNA Pathway Primed by Individual Transposons Is Linked to De Novo DNA Methylation in Mice. *Molecular Cell*, 2008.
- [82] X. Ding, H. Guan, and H. Li. Characterization of a piRNA binding protein Miwi in mouse oocytes. *Theriogenology*, 2013.

- [83] O. H. Tam *et al.* Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, 2008.
- [84] T. Watanabe *et al.* Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, 2008.
- [85] A. Aravin *et al.* A novel class of small RNAs bind to MILI protein in mouse testes. *Nature*, 2006.
- [86] A. Aravin *et al.* Developmentally Regulated piRNA Clusters Implicate MILI in Transposon Control. *Science*, 2007.
- [87] A. Girard *et al.* A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, 2006.
- [88] S. T. Grivna *et al.* A novel class of small RNAs in mouse spermatogenic cells. *Genes and Development*, 2006.
- [89] S. Ro *et al.* Cloning and expression profiling of testis-expressed piRNA-like RNAs. *RNA*, 2007.
- [90] A. Le Thomas *et al.* Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes and Development*, 2013.
- [91] S. Kuramochi-Miyagawa *et al.* DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Genes and Development*, 2008.
- [92] B. Czech and G. J. Hannon. One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. *Trends in Biochemical Sciences*, 2016.
- [93] R. Cordaux and M. A. Batzer. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 2009.
- [94] R. Rebollo, S. Farivar, and D. L. Mager. C-GATE - catalogue of genes affected by transposable elements. *Mobile DNA*, 2012.
- [95] B. G. Thornburg, V. Gotea, and W. Makalowski. Transposable elements as a significant source of transcription regulating signals. *Gene*, 2006.
- [96] G. Kunarso *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics*, 2010.
- [97] U. Beyer *et al.* Endogenous retrovirus drives hitherto unknown proapoptotic p63 isoforms in the male germ line of humans and great apes. *PNAS*, 2011.
- [98] D. C. Dolinoy. The agouti mouse model: an epigenetic biosensor for nutritional and environmental alterations on the fetal epigenome. *Nutrition Reviews*, 2008.
- [99] E. B. Chuong, N. C. Elde, and C. Feschotte. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science*, 2016.
- [100] V. J. Lynch *et al.* Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature Genetics*, 2011.
- [101] R. Sorek, G. Ast, and D. Graur. *Alu*-Containing Exons are Alternatively Spliced. *Genome Research*, 2002.
- [102] G. Lev-Maor *et al.* The Birth of an Alternatively Spliced Exon: 3' Splice-Site Selection in *Alu* Exons. *Science*, 2003.
- [103] N. Sela *et al.* Comparative analysis of transposed element insertion within human and mouse genomes reveals *Alu*'s unique role in shaping the human transcriptome. *Genome Biology*, 2007.
- [104] S. Shen *et al.* Widespread establishment and regulatory impact of *Alu* exons in human genes. *PNAS*, 2011.

- [105] L. Lin *et al.* Diverse Splicing Patterns of Exonized Alu Elements in Human Tissues. *PLoS Genetics*, 2008.
- [106] W. Tang *et al.* Secreted and membrane attractin result from alternative splicing of the human ATRN gene. *PNAS*, 2000.
- [107] S. J. Wheelan *et al.* Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. *Genome Research*, 2005.
- [108] J. L. Goodier, E. M. Ostertag, and H. H. Kazazian Jr. Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Human Molecular Genetics*, 2000.
- [109] C. M. Macfarlane *et al.* Transduction-Specific ATLAS Reveals a Cohort of Highly Active L1 Retrotransposons in Human Populations. *Human Mutation*, 2013.
- [110] J. V. Moran, R. J. DeBerardinis, and H. H. Kazazian Jr. Exon shuffling by L1 retrotransposition. *Science*, 1999.
- [111] O. K. Pickeral *et al.* Frequent Human Genomic DNA Transduction Driven by LINE-1 Retrotransposition. *Genome Research*, 2000.
- [112] J. M. C. Tubio *et al.* Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science*, 2014.
- [113] Y. Zhang *et al.* Repetitive elements and enforced transcriptional repression co-operate to enhance DNA methylation spreading into a promoter CpG-island. *Nucleic Acids Research*, 2012.
- [114] M. R. H. Est cio *et al.* SINE Retrotransposons Cause Epigenetic Reprogramming of Adjacent Gene Promoters. *Molecular Cancer Research*, 2012.
- [115] M. Friedli *et al.* Loss of transcriptional control over endogenous retroelements during reprogramming to pluripotency. *Genome Research*, 2014.
- [116] M. Friedli *et al.* Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Genome Research*, 2014.
- [117] A. Fort *et al.* Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nature Genetics*, 2014.
- [118] N. V. Fuchs *et al.* Human endogenous retrovirus K (HML-2) RNA and protein expression is a marker for human embryonic and induced pluripotent stem cells. *Retrovirology*, 2013.
- [119] A. E. Peaston *et al.* Retrotransposons Regulate Host Genes in Mouse Oocytes and Preimplantation Embryos. *Developmental Cell*, 2004.
- [120] X. Lu *et al.* The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nature Structural and Molecular Biology*, 2014.
- [121] T. S. Macfarlan *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, 2012.
- [122] M. Ohnuki *et al.* Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *PNAS*, 2014.
- [123] F. A. Santoni, J. Guerra, and J. Luban. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology*, 2012.
- [124] D. Kigami *et al.* MuERV-L Is One of the Earliest Transcribed Genes in Mouse One-Cell Embryos. *Biology of Reproduction*, 2003.
- [125] J. G ke *et al.* Dynamic Transcription of Distinct Classes of Endogenous Retroviral Elements Marks Specific Populations of Early Human Embryonic Cells. *Cell Stem Cell*, 2015.

- [126] J. Wang *et al.* Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*, 2014.
- [127] G. J. Faulkner *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics*, 2009.
- [128] A. Kapusta *et al.* Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLoS Genetics*, 2013.
- [129] P. Deininger *et al.* A comprehensive approach to expression of L1 loci. *Nucleic Acids Research*, 2017.
- [130] R. Johnson and R. Guigó. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA*, 2014.
- [131] P. E. Jacques, J. Jeyakani, and G. Bourque. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genetics*, 2013.
- [132] C. Lavialle *et al.* Paleovirology of ‘syncytins’, retroviral *env* genes exapted for a role in placentation. *Philosophical Transactions of the Royal Society B*, 2013.
- [133] S. M. Rawn and J. C. Cross. The Evolution, Regulation, and Function of Placenta-Specific Genes. *Annual Review of Cell and Developmental Biology*, 2008.
- [134] J. Denner. Expression and function of endogenous retroviruses in the placenta. *APMIS*, 2016.
- [135] G. J. Faulkner and J. L. Garcia-Perez. L1 Mosaicism in Mammals: Extent, Effects, and Evolution. *Trends in Genetics*, 2017.
- [136] M. Lynch. *The Origins of Genome Architecture*. Sinauer Associates, Inc., 2007.
- [137] H. Kaessmann, N. Vinckenbosch, and M. Long. RNA-based gene duplication: mechanistic and evolutionary insights. *Nature Reviews Genetics*, 2009.
- [138] J. Maestre *et al.* mRNA retroposition in human cells: processed pseudogene formation. *The EMBO Journal*, 1995.
- [139] O. Dhellin, J. Maestre, and T. Heidmann. Functional differences between the human LINE retrotransposon and retroviral reverse transcriptases for in vivo mRNA reverse transcription. *The EMBO Journal*, 1997.
- [140] M. R. Kubiak and I. Makalowska. Protein-Coding Genes’ Retrocopies and Their Functions. *Viruses*, 2017.
- [141] F. N. Carelli *et al.* The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome Research*, 2016.
- [142] S. Tan *et al.* LTR-mediated retroposition as a mechanism of RNA-based duplication in metazoans. *Genome Research*, 2016.
- [143] Z. Zhang *et al.* Millions of Years of Evolution Preserved: A Comprehensive Catalog of the Processed Pseudogenes in the Human Genome. *Genome Research*, 2003.
- [144] Z. Zhang, N. Carriero, and M. Gerstein. Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends in Genetics*, 2004.
- [145] D. Pain *et al.* Multiple retropseudogenes from pluripotent cell-specific gene expression indicates a potential signature for novel gene identification. *Journal of Biological Chemistry*, 2005.
- [146] A. C. Marques *et al.* Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. *Genome Biology*, 2008.
- [147] F. Burki and H. Kaessmann. Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nature Genetics*, 2004.

- [148] L. Rosso *et al.* Mitochondrial Targeting Adaptation of the Hominoid-Specific Glutamate Dehydrogenase Driven by Positive Darwinian Selection. *PLoS Genetics*, 2008.
- [149] V. Mastorodemos *et al.* Molecular basis of human glutamate dehydrogenase regulation under changing energy demands. *Journal of Neuroscience Research*, 2004.
- [150] D. M. Sayah *et al.* Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature*, 2004.
- [151] G. Brennan, Y. Kozyrev, and S. L. Hu. TRIMCyp expression in Old World primates *Macaca nemestrina* and *Macaca fascicularis*. *PNAS*, 2008.
- [152] C. A. Virgen *et al.* Independent genesis of chimeric TRIM5-cyclophilin proteins in two primate species. *PNAS*, 2008.
- [153] N. Vinckenbosch *et al.* Evolutionary fate of retroposed gene copies in the human genome. *PNAS*, 2006.
- [154] A. C. Marques *et al.* Emergence of Young Human Genes after a Burst of Retroposition in Primates. *PLoS Biology*, 2005.
- [155] E. Betrán, K. Thornton, and M. Long. Retroposed New Genes Out of the X in *Drosophila*. *Genome Research*, 2002.
- [156] L. Potrzebowski *et al.* Chromosomal Gene Movements Reflect the Recent Origin and Biology of Therian Sex Chromosomes. *PLoS Biology*, 2008.
- [157] J. J. Emerson *et al.* Extensive gene traffic on the mammalian X chromosome. *Science*, 2004.
- [158] P. J. Wang. X chromosomes, retrogenes and their role in male reproduction. *Trends in Endocrinology and Metabolism*, 2004.
- [159] E. M. Muro and M. A. Andrade-Navarro. Pseudogenes as an alternative source of natural antisense transcripts. *BMC Evolutionary Biology*, 2010.
- [160] O. Bryzghalov, M. W. Szcześniak, and I. Makalowska. Retroposition as a source of antisense long non-coding RNAs with possible regulatory functions. *Acta Biochimica Polonica*, 2016.
- [161] M. J. Milligan *et al.* Global Intersection of Long Non-Coding RNAs with Processed and Unprocessed Pseudogenes in the Human Genome. *Frontiers in Genetics*, 2016.
- [162] P. Johnsson *et al.* A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. *Nature Structural and Molecular Biology*, 2013.
- [163] J. L. Rinn *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 2007.
- [164] M. R. Fabian, N. Sonenberg, and W. Filipowicz. Regulation of mRNA Translation and Stability by microRNAs. *Annual Review of Biochemistry*, 2010.
- [165] L. Poliseno *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, 2010.
- [166] G. Yu *et al.* Pseudogene PTENP1 Functions as a Competing Endogenous RNA to Suppress Clear-Cell Renal Cell Carcinoma Progression. *Molecular Cancer Therapeutics*, 2014.
- [167] C. L. Chen *et al.* Suppression of hepatocellular carcinoma by baculovirus-mediated expression of long non-coding RNA PTENP1 and MicroRNA regulation. *Biomaterials*, 2015.
- [168] L. Wang *et al.* Pseudogene OCT4-pg4 functions as a natural micro RNA sponge to regulate OCT4 expression by competing for miR-145 in hepatocellular carcinoma. *Carcinogenesis*, 2013.

- [169] H. Peng *et al.* Pseudogene INTS6P1 regulates its cognate gene INTS6 through competitive binding of miR-17-5p in hepatocellular carcinoma. *Oncotarget*, 2015.
- [170] S. Katayama *et al.* Antisense Transcription in the Mammalian Transcriptome. *Science*, 2005.
- [171] K. V. Morris and J. S. Mattick. The rise of regulatory RNA. *Nature Reviews Genetics*, 2014.
- [172] S. A. Korneev, J.-H. Park, and M. O'Shea. Neuronal Expression of Neural Nitric Oxide Synthase (nNOS) Protein Is Suppressed by an Antisense RNA Transcribed from an NOS Pseudogene. *Journal of Neuroscience*, 1999.
- [173] E. J. Devor. Primate MicroRNAs *miR-220* and *miR-492* Lie within Processed Pseudogenes. *Journal of Heredity*, 2006.
- [174] T. Hirano *et al.* Small RNA profiling and characterization of piRNA clusters in the adult testes of the common marmoset, a model primate. *RNA*, 2014.
- [175] L. Pantano *et al.* The small RNA content of human sperm reveals pseudogene-derived piRNAs complementary to protein-coding genes. *RNA*, 2015.
- [176] T. Watanabe *et al.* Retrotransposons and pseudogenes regulate mRNAs and lncRNAs via the piRNA pathway in the germline. *Genome Research*, 2015.
- [177] D. Gebert *et al.* piRNAs from Pig Testis Provide Evidence for a Conserved Role of the Piwi Pathway in Post-Transcriptional Gene Regulation in Mammals. *PLoS One*, 2015.
- [178] T. Watanabe *et al.* Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, 2008.
- [179] W.-L. Chan *et al.* Transcribed pseudogene ψ PPM1K generates endogenous siRNA to suppress oncogenic cell growth in hepatocellular carcinoma. *Nucleic Acids Research*, 2013.
- [180] University of Oregon Cresko Lab. RNA-seqlopedia. <https://rnaseq.uoregon.edu/>.
- [181] S. Marguerat and J. Bähler. RNA-seq: from technology to biology. *Cellular and Molecular Life Sciences*, 2010.
- [182] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 2009.
- [183] A. Oshlack, M. D. Robinson, and M. D. Young. From RNA-seq reads to differential expression results. *Genome Biology*, 2010.
- [184] N. L. Bray *et al.* Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 2016.
- [185] R. Patro, S. M. Mount, and C. Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, 2014.
- [186] R. Patro *et al.* Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 2017.
- [187] T. J. Treangen and S. L. Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 2012.
- [188] J. C. Marioni *et al.* RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 2008.
- [189] U. Nagalakshmi *et al.* The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, 2008.

- [190] A. Mortazavi *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 2008.
- [191] B. Li *et al.* RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 2010.
- [192] B. Li and C. N. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 2011.
- [193] T. M. Keane *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 2011.
- [194] K. D. Hansen, B. Langmead, and R. A. Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 2012.
- [195] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 2010.
- [196] T. Gingeras. ENCSR000AJU. <https://www.encodeproject.org/experiments/ENCSR000AJU/>, 2014.
- [197] Wellcome Trust Sanger Institute. Mouse Genomes Project. <http://www.sanger.ac.uk/science/data/mouse-genomes-project>.
- [198] A. Yates *et al.* Ensembl 2016. *Nucleic Acids Research*, 2016.
- [199] Ensembl release 84. <http://mar2016.archive.ensembl.org/index.html>.
- [200] C. Tyner *et al.* The UCSC Genome Browser database: 2017 update. *Nucleic Acids Research*, 2017.
- [201] A. F. A. Smit, R. Hubley, and P. Green. RepeatMasker Open-3.0., 1996-2010.
- [202] W. Bao, K. K. Kojima, and O. Kohany. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 2015.
- [203] R. Hubley *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Research*, 2016.
- [204] O. Kohany *et al.* Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, 2006.
- [205] D. R. Hoen *et al.* A call for benchmarking transposable element annotation methods. *Mobile DNA*, 2015.
- [206] I. R. Arkhipova. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mobile DNA*, 2017.
- [207] R. Baertsch *et al.* Retrocopy contributions to the evolution of the human genome. *BMC Genomics*, 2008.
- [208] N. A. O’Leary *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 2016.
- [209] S. Andrews. FastQC: A quality control tool for high throughput sequence data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- [210] M. Martin. Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads. *EMBnet.journal*, 2011.
- [211] C. Del Fabbro *et al.* An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLoS One*, 2013.
- [212] A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 2014.

- [213] M. MacManes. On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics*, 2014.
- [214] P. G. Engström *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods*, 2013.
- [215] A. Dobin *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 2013.
- [216] T. D. Wu *et al.* GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. *Methods in Molecular Biology*, 2016.
- [217] G. R. Grant *et al.* Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, 2011.
- [218] D. Kim, B. Langmead, and S. L. Salzberg. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 2015.
- [219] S. M. E. Sahraeian *et al.* Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nature Communications*, 2017.
- [220] L. Wang, S. Wang, and W. Li. RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, 2012.
- [221] M. Guttman *et al.* *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, 2010.
- [222] C. Trapnell *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 2010.
- [223] M. Pertea *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 2015.
- [224] C. Zhang *et al.* Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics*, 2017.
- [225] [http://rseqc.sourceforge.net/#spilt-bam-py\[sic\]](http://rseqc.sourceforge.net/#spilt-bam-py[sic]).
- [226] H. Li *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 2009.
- [227] J. Fu *et al.* ballgown: Flexible, isoform-level differential expression analysis. *Nature Protocols*, 2016.
- [228] M. Pertea *et al.* Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Bioconductor*, 2017.
- [229] C. R. Williams *et al.* Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. *BMC Bioinformatics*, 2017.
- [230] S. F. Altschul *et al.* Basic local alignment search tool. *Journal of Molecular Biology*, 1990.
- [231] BLAST+ Suite. https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download.
- [232] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 1987.
- [233] D. Zwillinger and S. Kokoska. *CRC Standard Probability and Statistics Tables and Formulae*. Chapman and Hall: New York, 2000.
- [234] K. Tretyakov. Area-weighted venn-diagrams for Python/matplotlib, 2012–.
- [235] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 1965.

- [236] R. D'Agostino and E. S. Pearson. An omnibus test of normality for moderate and large sample size. *Biometrika*, 1973.
- [237] EMBOSS. <http://emboss.sourceforge.net/>.
- [238] M. Krzywinski *et al.* Circos: an Information Aesthetic for Comparative Genomics. *Genome Research*, 2009.
- [239] O. Tange. GNU Parallel - The Command-Line Power Tool. ;login: *The USENIX Magazine*, 2011.
- [240] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering*, 2007.
- [241] S. van der Walt, S. C. Colbert, and G. Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing In Science & Engineering*, 2011.
- [242] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–.
- [243] F. Pedregosa *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2011.
- [244] R. A. Elbarbary, B. A. Lucas, and L. E. Maquat. Retrotransposons as regulators of gene expression. *Science*, 2016.
- [245] T. Penzkofer *et al.* L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. *Nucleic Acids Research*, 2017.
- [246] J. F. Kugel and J. A. Goodrich. Non-coding RNAs: key regulators of mammalian transcription. *Cell*, 2012.
- [247] S. Geisler and J. Collier. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nature Reviews Molecular Cell Biology*, 2013.
- [248] J. H. Noh *et al.* Cytoplasmic functions of long noncoding RNAs. *Wires RNA*, 2018.
- [249] The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Research*, 2015.
- [250] J. F. Hughes and D. C. Page. The Biology and Evolution of Mammalian Y Chromosomes. *Annual Review of Genetics*, 2015.
- [251] D. Pan and L. Zhang. Burst of young retrogenes and independent retrogene formation in mammals. *PLoS One*, 2009.
- [252] IGKV Gene Cluster. <http://www.genecards.org/cgi-bin/carddisp.pl?gene=IGKV0>.
- [253] F. W. Scholz and M. A. Stephens. K-Sample Anderson-Darling Tests. *Journal of the American Statistical Association*, 1987.
- [254] S. Kannan *et al.* Transposable element insertions in long intergenic non-coding RNA genes. *Front. Bioeng. Biotechnol.*, 2015.
- [255] M. E. Dinger *et al.* NRED: a database of long noncoding RNA expression. *Nucleic Acids Research*, 2009.
- [256] C. Gong and L. E. Maquat. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature*, 2011.
- [257] B. A. Lucas *et al.* Evidence for convergent evolution of SINE-directed Staufen-mediated mRNA decay. *PNAS*, 2018.
- [258] A. Wenzel, E. Akbasli, and J. Gorodkin. RIssearch: fast RNA-RNA interaction search using a simplified nearest-neighbor energy model. *Bioinformatics*, 2012.
- [259] G. Terai *et al.* Comprehensive prediction of lncRNA-RNA interactions in human transcriptome. *BMC Genomics*, 2016.

- [260] M. W. Szczepaniak and I. Makowska. lncRNA-RNA Interactions across the Human Transcriptome. *PLoS One*, 2016.
- [261] Y. Yano *et al.* A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. *Journal of Molecular Medicine*, 2004.
- [262] P. M. Harrison *et al.* Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Research*, 2005.
- [263] M. Sun *et al.* SAGE is far more sensitive than EST for detecting low-abundance transcripts. *BMC Genomics*, 2004.
- [264] J. R. McCarrey and A. D. Riggs. Determinator-inhibitor pairs as a mechanism for threshold setting in development: a possible function for pseudogenes. *PNAS*, 1986.
- [265] P. G. Hawkins and K. V. Morris. Transcriptional regulation of Oct4 by a long non-coding RNA antisense to Oct4-pseudogene 5. *Transcription*, 2010.
- [266] A. R. Bassett *et al.* Considerations when investigating lncRNA function *in vivo*. *eLife*, 2014.
- [267] J. L. Rinn and H. Y. Chang. Genome Regulation by Long Noncoding RNAs. *Annual Review of Biochemistry*, 2012.
- [268] SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology*, 2014.
- [269] J. J  nes *et al.* A comparative study of RNA-seq analysis strategies. *Briefings in Bioinformatics*, 2015.
- [270] N. F. Lahens *et al.* IVT-seq reveals extreme bias in RNA-sequencing. *Genome Biology*, 2014.
- [271] N. Maeda *et al.* Transcript Annotation in FANTOM3: Mouse Gene Catalog Based on Physical cDNAs. *PLoS Genetics*, 2006.
- [272] A. Sookdeo *et al.* Revisiting the evolution of mouse LINE-1 in the genomic era. *Mobile DNA*, 2013.
- [273] H. Khan, A. Smith, and S. Boissinot. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Research*, 2006.
- [274] N. B. Adey *et al.* Rodent L1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences. *Molecular Biology and Evolution*, 1994.
- [275] I. M. Serdobova and D. A. Kramerov. Short Retroposons of the B2 Superfamily: Evolution and Application for the Study of Rodent Phylogeny. *Journal of Molecular Evolution*, 1998.
- [276] E. Lee *et al.* Landscape of Somatic Retrotransposition in Human Cancers. *Science*, 2012.
- [277] S. Solyom *et al.* Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Research*, 2012.
- [278] D. Jia *et al.* Structure of Dnmt3a bound to Dnmt3L suggests a model for *de novo* DNA methylation. *Nature*, 2007.
- [279] M. Okano *et al.* DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell*, 1999.
- [280] E. Li, T. H. Bestor, and R. Jaenisch. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, 1992.
- [281] J. H. Crichton *et al.* Defending the genome from the enemy within: mechanisms of retrotransposon suppression in the mouse germline. *Cellular and Molecular Life Sciences*, 2014.

- [282] P. Yang, Y. Wang, and T. S. Macfarlan. The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. *Trends in Genetics*, 2017.
- [283] G. Ecco, M. Imbeault, and D. Trono. KRAB zinc finger proteins. *Development*, 2017.
- [284] M. Imbeault, P.-Y. Helleboid, and D. Trono. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, 2017.
- [285] F. M. J. Jacobs *et al.* An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature*, 2014.
- [286] G. Wolf *et al.* The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. *Genes and Development*, 2015.
- [287] S. Iyengar and P. J. Farnham. KAP1 Protein: An Enigmatic Master Regulator of the Genome. *The Journal of Biological Chemistry*, 2011.
- [288] D. C. Leung and M. C. Lorincz. Silencing of endogenous retroviruses: when and why do histone marks predominate? *Trends in Biochemical Sciences*, 2012.
- [289] A. K. Lim *et al.* The nuage mediates retrotransposon silencing in mouse primordial ovarian follicles. *Development*, 2013.
- [290] D. Banville and Y. Boie. Retroviral long terminal repeat is the promoter of the gene encoding the tumor-associated calcium-binding protein oncomodulin in the rat. *Journal of Molecular Biology*, 1989.
- [291] K. L. Yap *et al.* Molecular Interplay of the Noncoding RNA ANRIL and Methylated Histone H3 Lysine 27 by Polycomb CBX7 in Transcriptional Silencing of INK4a. *Molecular Cell*, 2010.